Ethics, Fitting Attitudes, and Practical Reason: A Theory of Normative Facts

by

Howard L. M. Nye

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2009

Doctoral Committee:

Professor Allan F. Gibbard, Co-Chair
Professor Peter A. Railton, Co-Chair
Professor William J. Gehring
Professor Stephen L. Darwall, Yale University

For

Large Twin, Number Two, Small Twin, and Potzie

In Loving Memory

Acknowledgements

The ideas that shaped this dissertation have been long in the making, and I have received help in developing them from far too many people to hope to remember or name. But I should thank first and foremost Stephen Darwall, Allan Gibbard, John Ku, and Peter Railton, who provided me with invaluable assistance when this material was in its formative stages, and who have been terrifically helpful in so many ways ever since.

I also owe a special debt of gratitude to those of my graduate student colleagues with whom I have had many very helpful discussions about this material and topics that bore directly upon it, including Michael Allers, Aaron Bronfman, Vanessa Carbonell, Ian Flora, Dustin Locke, Erica Stonestreet, Ivan Mayerhoffer, Neil Mehta, David Plunkett, and Neil Sinhababu.

I would like to thank everyone who has helped me in conversations about the ideas presented in this dissertation and issues related to them. While I could never list every such person by name, among many others I am grateful to Elizabeth Anderson, David Braddon-Mitchell, Alex Byrne, Jonathan Dancy, Justin D'Arms, Joshua Dever, Andy Egan, William Gehring, Richard Holton, Frank Jackson, Daniel Jacobson, Jim Joyce, Philip Kitcher, Boris Kment, Rae Langton, Louis Loeb, Eric Lormand, Alastair Norcross, Douglas Portmore, Jacob Ross, Geoffrey Sayre-McCord, Chandra Sirpada, Brad Skow, Jamie Tappenden, and Ralph Wedgewood. I am also indebted to audiences from several workshops, conferences, and talks at which I presented parts of this material, including those at the University of Michigan, the University of Pittsburg, Arizona State University, the University of Florida, Oxford University, MIT, Carleton College, and the University of Alberta.

I would like to thank my parents, Robert Nye and Rebecca Morrow-Nye for their tremendous love and support. Finally, I am grateful to my cats, Kipling, Emily, and Zapata, as well as to my mice, Large Twin, Number Two, Small Twin, and Potzie, for their constant affection and companionship.

Table of Contents

List of Figures

Chapter 1

Puzzles About Ethics and Normativity

1.1. Ethical Judgments

Ethical judgments play a central role in our decisions about what to do. We take ourselves to be constrained by the need to avoid doing what is morally wrong, and we often strive to do what would be morally good or virtuous. In everyday life, we pass up opportunities to lie, cheat, or steal from strangers on the grounds that it would be wrong, and we extend extra help to people because we take it to be a good thing to do. In making more momentous decisions about what kinds of things to do with our lives, our views about what we owe to others and which activities are worthy often lead us to take up volunteer work and occupations that are more stressful or lower paying than we might otherwise like. Our views about the extent to which various people are morally blameworthy for what they have done also influence our choices, from decisions about whether to scold a family member to decisions about how our societies should allocate punishments.

The term 'ethics' is sometimes used in a narrow sense that makes it virtually synonymous with 'morality', but philosophers often use the term in a wider sense that encompasses evaluative considerations that are not essentially moral. In making decisions we think about the outcomes of the various things we could do, and we think about which of these would be good, which would be bad, and which would be better than others. Central to our thinking about the goodness of outcomes are our judgments about the welfare of those beings we care about – judgments, that is, about what would be good or bad for them, and what will make them better or worse off. While these judgments about what harms or benefits a being are not themselves about morality's

demands or recommendations, they do play an absolutely central role in our moral thinking about what would be wrong or good to do.

While morality, the value of outcomes, and welfare are the broadly ethical categories that are perhaps most often discussed by contemporary philosophers, there are other judgments that I think are important for a general understanding of the relationship between ethical thought and practical reasoning, or thinking about what to do. First, we seem to think that certain things are intrinsically valuable in ways that are not a matter of their making sentient beings better off. We often seem to think that things like beautiful artworks and natural wonders are valuable quite independently of their contribution to how well off anyone is. This seems, moreover, to have an effect on our choices - direct valuing of the arts and the natural environment may well stand behind quite a few of our personal and policy decisions. Our decisions about how to spend our time are also influenced by our judgments about the intrinsic value of certain activities like playing music and studying mathematics. While we certainly tend to think that such activities make our lives worthwhile, we often seem to think that our reasons to engage in them have to do with their possessing a kind of value that is quite independent of what they do for us.

Second, there seems to be a set of ethical concepts that are structurally similar to our moral concepts but that play a somewhat different role in our lives. We often tend to think that things like lazy or cowardly behavior are *shameful* or *lowly* quite apart from any moral objections we might have to them. In contexts in which acts of laziness or cowardice do no harm to others, we tend to feel scorn or disdain for those who engage in them even if we do not feel angry or morally indignant with them. Similarly, we tend to think that things like intellectual or athletic accomplishments and exemplary exercises of hard work or discipline are *excellent* or virtuous in a way that is distinct from being morally good or virtuous. When an achievement is made or hard work and discipline is exercised in pursuits that are largely unconcerned with the interests of others, we tend to have a kind of esteem for it that is somewhat distinct from that which we have towards people who sacrifice a great deal to help others.

Finally, there are a variety of other evaluative judgments that I think play a role in our practical reasoning that is similar to that played by the foregoing ethical and

evaluative judgments. These include judgments about which events are tragic or grievous, which features of others are envious, and which entities and situations are dangerous.

For ease of exposition (and for reasons that will become clearer later on) I shall refer to all of the foregoing kinds of judgments as 'ethical judgments'. What these judgments seem to have in common is that they entail judgments about the kinds of pro- and con-attitudes we should take towards various things. An act's moral wrongness, for instance, entails that we should have a particular kind of con-attitude towards it – namely that we should feel obligated not to perform the act, which involves a kind of motivation not to perform it. Similarly, the goodness of an outcome or state of affairs entails that we should have certain pro-attitudes towards that state of affairs; that we should be glad of its having come about, wish that it would come about, or desire that does come about. These kinds of attitudes towards a state of affairs involve motivations to do what we can to maintain it or bring it about. Likewise, a person's status as morally good or excellently virtuous entails that we should take certain pro-attitudes towards her, namely that we should feel a kind of esteem for her, which is tinged with something like gratitude if she has done good, and tinged with something like awe if she has done well. These attitudes of esteem involve motivation to do what we can to emulate the person esteemed if placed in similar circumstances.

The pro- and con- attitudes we take to be justified in making ethical judgments thus include those we commonly refer to as emotions, desires, and aversions. These attitudes involve particular motivations or tendencies towards motivation, and have a kind of phenomenology or way it is like to subjectively experience having them. The pro-attitudes involve what are intuitively "positive" motivations and feelings that attract one to the act, outcome, or person towards which they are felt. The con-attitudes involve intuitively "negative" motivations and feelings that repel one from the act, outcome, or person toward which they are felt. I shall refer to these pro- and con-attitudes collectively as 'valenced attitudes'.

In this dissertation I will be offering an account of what we are doing when we make ethical judgments. My theory will seek to explain what makes our ethical

judgments true and what ethics has to do with reasons for action. The theory I develop will be a theory not only about ethics but about what philosophers call "normativity" more generally. It will be a theory about what makes an action, belief, or attitude something that we should do, think, or feel. In this chapter I outline some of the main things that seem puzzling about ethics and normativity, and I briefly sketch how my theory will seek to solve the puzzles.

1.2. Normative Judgments

Along with practical judgments about what we should do and epistemic judgments about what we should believe, judgments about what valenced attitudes we should have are paradigmatic instances of those judgments philosophers call 'normative'. Indeed, we may understand a normative judgment as one that entails a judgment that someone should or ought to have a certain response (e.g. do, think, or feel something), and we can understand ethical judgments as normative because they entail judgments about what valenced attitudes we ought to have.[1] Normative or "prescriptive" judgments are usually contrasted with merely descriptive judgments, and I think that this contrast is made for very good reason. Thinking that you should have a certain response seems to be importantly different from thinking that the response has the kinds of features that we cite in causal and mathematical explanations of the way the world is. Perhaps most obviously, our judgments about what we should do, think, and feel play a role in guiding

[1] I should clarify that when I say that a normative judgment entails a judgment about how to respond, I mean that the normative judgment's truth guarantees the truth of some judgment about how we should respond in virtue of meaning and deductive logic alone. Thus, while I can correctly infer from the fact that going to the store will be painful that I have some reason not to go to the store, the judgment that going to the store will be painful is not itself normative because it does not guarantee in virtue of meaning and logic alone that I have reason to respond in some way. (To make my inference deductively valid we need the premise that I have reason to avoid pain, which is no logical or analytic truth but a substantive normative fact.)

A genuine problem for understanding normative judgments as those that entail judgments about how to respond is that any judgment *J* will entail the disjunction of *J* with any judgment about how anyone should respond that you like. Perhaps we should try saying that an atomic normative judgment is one that entails an atomic judgment about how to respond, and a normative judgment is either an atomic normative judgment or a judgment that is built up from a set of judgments that includes an atomic normative judgment.

our actions, beliefs, and valenced attitudes in a way that our ordinary descriptive beliefs do not.

We have a significant tendency to do, think, and feel what we think we should. Of course, our responses do not always conform to our judgments about what they should be. You can think that you really ought to get out of bed when the alarm clock rings, but fail to do so nonetheless. You can also think that you have overwhelming reason to believe that the many worlds interpretation of Quantum Mechanics correctly describes reality, but find yourself unable to believe (or have a very high degree of belief in) the theory. Similarly, you can in the depths of depression continue to think that knowledge is good and something you should want, but find yourself with no actual desire for it or motivation to pursue it. But these kinds of examples of the failures of our actions, beliefs, and desires to respond to our normative judgments owe much of their salience to the fact that they are exceptions. By and large, our judgments about what to do, believe, and feel play a role in explaining our actual actions, beliefs, and valenced attitudes; our responses would be different if we were to hold different views about which responses we should have. Our judgments about what we should do, believe, and feel thus seem to exert a kind of causal pressure on our actions and attitudes; they have a propensity to cause us to have the responses that we think we should have. This propensity remains even if, due to the presence of inclinations or tendencies to the contrary, it fails to determine what we actually do, think, and feel.

Of course, against the right background of beliefs and desires, our ordinary descriptive judgments will have a propensity to influence our actions, beliefs, and valenced attitudes. If you happen to want to buy a cotton jacket, then coming to judge that a particular jacket is made of cotton can cause you to buy it. If you happen to believe that drinking leads to dancing, then coming to believe that there is drinking in the next room can cause you to believe that there will soon be dancing in there. And if you happen to want the potholes to be filled, then coming to believe that candidate *X* will fill the potholes if elected can cause you to desire the election of candidate *X*. But normative judgments about what we should do, think, and feel seem to have a different kind of influence on our responses. This kind of influence is less contingent than that of our

descriptive judgments and it does not depend upon our background beliefs and desires in the same kind of way.

Coming to judge that you have reason to buy a particular jacket seems capable of causing you to buy it in the absence of a contingent desire to do whatever you have reason to do. Similarly, coming to judge that your evidence supports the view that there will be dancing in the next room seems capable of causing you to believe that there will be in the absence of a contingent belief that whatever your evidence supports is true. And coming to judge that it would be good or desirable for the potholes to be filled seems capable of causing you to want them to be filled in the absence of a contingent background desire for whatever states of affairs are good or desirable. Rather than requiring beliefs and desires that we could have lacked, the influence of our normative judgments about how we should respond on how we do respond seems essential to these judgments. We simply would not count as making a judgment about what to do, think, or feel if we failed to be in a state that exerted causal pressure on our actions, beliefs, and valenced attitudes.

Normative judgments are thus unlike ordinary descriptive judgments in that they seem to have an essential propensity to influence our responses. But there are other respects in which our normative judgments are very much like descriptive judgments. Most notably, we can wonder whether our judgments about what to do, believe, and feel are true, and we have procedures for determining whether or not they are. In practical, epistemic, and ethical reflection, we inquire into things like whether we really should pursue knowledge for its own sake, whether our evidence supports believing in God, and whether killing early fetuses is wrong or something we should feel obligated not to do. While this kind of normative inquiry is pursued by professional philosophers, it is actually a mainstay of ordinary life. It is not merely that almost all ordinary people think about such momentous questions as the foregoing and that the lives of many if not most are deeply shaped by their views about how to answer them. It is also that everyday life confronts us with more mundane but no less theoretically important normative questions, including those about whether the health benefits of additional exercise are really worth the time it will take, the extent to which our evidence supports the hypothesis that Joe

stole the cookie from the cookie jar, and whether we should really feel obligated to do someone the favor she asks or invite her to a party in light of our history of interaction with her.

Certain kinds of normative questions might be settled by settling ordinary descriptive questions. Two people might agree that it is wrong to kill early fetuses for convenience just in case such fetuses are sentient, but disagree about whether it is actually wrong to kill them because they disagree about whether early fetuses are sentient. Lying behind this sort of normative disagreement is an important kind of normative agreement about the conditions under which an act is wrong. It is helpful here to distinguish between two kinds of normative questions. *Basic normative questions* are questions about what to do, think, or feel in a circumstance the merely descriptive features of which are simply stipulated or assumed. Thus whether it is wrong to kill fetuses assuming they are conscious (and other merely descriptive features of them) is a basic normative question, to which the above parties both give an affirmative answer. On the other hand, *applied normative questions* are questions about what to do, think, and feel in a circumstance the descriptive features of which are not simply stipulated. These applied questions will often ask what we should do, think, and feel in our actual circumstances, but we can also ask such questions about counter-factual scenarios as "what should Kennedy have done had the Bay of Pigs Invasion succeeded?" and "what should we have concluded about the theory of evolution had there been no fossil record?" We can think of our answers to applied normative questions as "factored" into an answer to a basic normative question about what to do, think, and feel in various stipulated circumstances, together with an answer to a merely descriptive question about which of those circumstances accurately describe those we are actually considering.

While some have been skeptical of the existence of such a thing as inquiry into basic normative questions (or at least of inquiry into basic questions about what to do and feel),[2] it seems clear that we engage in it all the time. Such basic normative inquiry proceeds by our soliciting and seeking out a reflective equilibrium among our various

[2] Early emotivists like Ayer (1936) and Stevenson (1944) are rather striking examples of such skeptics.

intuitions about what we should do, believe, and feel.[3] While talk of "intuitions" sometimes suggests an exclusive focus on intuitions about how to respond in particular cases, I intend my use of the phrase 'normative intuition' to refer to any spontaneous, un-inferred appearance about what we should do, think, or feel, no matter how general its content. Thus, we have intuitions about the general kinds of things that we have reason to do, think, or feel, as when it immediately seems to be true that we have reason to avoid our own pain, believe our sensory evidence, and feel guilty for harming others. We also have intuitions about the kinds of features of a circumstance that do or do not make a difference to what we should do, believe, or feel in it, as when it immediately seems that someone's bare physical distance from us is irrelevant to how much we should care about or help her, or that the sheer number of people who believe something is irrelevant to whether the belief is justified. We even have intuitions about the plausibility of general principles concerning what we should do, think, and feel, as when it just seems immediately compelling that we should promote the pleasure and prevent the pain of all sentient creatures, or that *ceteris paribus* we should believe the simplest theory consistent with our evidence.

Reflective equilibrium methods of normative inquiry inevitably involve discounting and debunking certain of our intuitions about what responses we should have, and trying to explain them away as the product of something other than an accurate perception of what we should in fact do, think, and feel. For instance, as Unger (1996) (following Singer (1972)) points out, our intuitions about how little we are obligated to sacrifice to help distant strangers and our intuitions about how much we are obligated to sacrifice to help nearby strangers seem to conflict with our intuitions about how little someone's bare physical distance matters to the strength of our reasons to help her. Unger argues convincingly that our intuitions about how little we owe to distant strangers are actually normative illusions that – unlike our intuitions about what we owe to nearby strangers and how little distance matters – have been produced by various cognitive biases rather than the truth about our reasons to help people.

[3] For characterizations of normative inquiry as seeking such a reflective equilibrium, see for instance (Goodman 1954), (Rawls 1971), (Daniels 1979), and (McMahan 2000).

Similarly, we might be inclined to think that, *ceteris paribus*, the greater the number of entities a theory posits, the less reason we have to believe the theory. But we might attend to the distinction between numbers of entities and numbers of kinds of entities. We might have the intuition that the fact that theory A posits only 10^{29} electrons while theory B posits 10^{37} electrons does not seem to count in favor of believing theory A over theory B.[4] We might also see that most of the theories we think we should believe on grounds of ontological simplicity are in fact theories that posit fewer kinds of entities rather than simply fewer entities. From this we might conclude that our intuitions that it counts against a theory for it to posit more entities independent of more kinds of entities is actually an erroneous result of mistaking the epistemic relevance of numbers of kinds of entities for the epistemic relevance of numbers of entities.

In the same way, a theist might find it intuitively compelling that one should only feel guilt for doing whatever violates God's commandments, and that one should feel guilt for doing such things simply because they violate God's commandments. But the theist might reflect upon hypothetical cases and have the intuitions that one should feel guilt for torturing children even in cases in which God has not forbidden it, and that one should not feel guilt for refusing to torture children even in cases in which God has commanded it. In light of this, the theist might conclude that her intuition about the plausibility of the principle requiring guilt only on account of violating God's commandments was actually mistaken. Perhaps it was the product of indoctrination, or something like a confusion between an act's having the feature of violating God's commandments and an act's having such features as harming children. Since the theist presumably takes these features to be correlated in the actual world, it would not be surprising that she would have confused which of them is of basic normative relevance.[5]

There is thus a very salient way in which our methods of basic normative inquiry seem to presuppose that our normative judgments are in fact descriptive judgments, or representations of the kinds of facts that can play a role in causal explanations of such

[4] See (Lewis 1973).

[5] Since reflective equilibrium methods as I characterize them involve the debunking of intuitions about cases and principles alike, they encompass methods that incorporate views about whether intuitions about cases tend to be more or less reliable than intuitions about general principles. Reflective equilibrium methods as I understand them thus include everything ranging from approaches like that advocated by Singer (1974), which are more skeptical of intuitions about particular cases, to approaches like that advocated by Kamm (1993), which are more skeptical of intuitions about general principles.

phenomena as why we hold the beliefs that we do.[6] When our normative intuitions conflict and we seek to explain some of them away, we seem to be trying to show that some of our intuitions have not – *whereas the rest of our intuitions have* – been generated by cognitive mechanisms that accurately detect facts about what we should actually do, think, and feel. We seem to assume that there is a kind of causal-explanatory relationship between facts about how we should respond and our intuitions about how we should respond, which has become interfered with or scrambled in the case of erroneous intuitions. In this sense we take our intuitions to have a basic (but defeasible) tendency to reflect normative reality in much the way we take our perceptual experiences to have a basic (but defeasible) tendency to reflect the reality of the external world.

More generally, whenever we inquire into whether some proposition *P* is true, we seem to be trying to align our judgments about *P*'s truth with the facts of the matter about whether *P* is true. Since we do not seek to be right by accident, we seem to be trying to achieve some kind of causal or explanatory contact between our judgment about *P* and the fact that *P* (if *P* is true) or the fact that not-*P* (if *P* is false). Another way to put this might be to say that when we inquire into whether *P* is true, we are attempting to arrive at knowledge that *P* or knowledge that not-*P*. But to know that *P* seems to require that *P*'s truth figures into the explanation of one's judgment that *P*.[7] As such, basic normative

[6] A role in the explanation of such ordinary descriptive phenomena such as our having the thoughts we do is exactly the dividing line drawn by (Gibbard 2003, 183-185) between descriptive truths and facts and the kind of normaitve truths and facts that an anti-descriptivist (of the so-called 'quasi-realist' variety) can claim to exist.

[7] I mean to construe "figuring into an explanation" so as to include both those facts that are indispensible to the explanation and those facts that are analytically entailed by those that are indispensible. It seems plausible to suppose that the explanation of a fact must entail that it (or – in indeterministic cases – the probability that it) obtains, and that when someone knows something about the future there must be a common explanation of both her belief and what she knows (see for instance (Hempel 1965), (Railton 1978), and (Dretske 1981)). If this is right then my kind of requirement on knowledge can, like Dretske's (1981) account of knowledge as information caused belief, subsume what seems right about the causal theory of knowledge but also capture the explanatory relationship that must obtain between beliefs about the future and their truth for them to constitute knowledge. Moreover, because a fact's causing a belief is presumably not the only way the fact can enter into the ontic explanation of - or reason why it is the case that - one holds the belief, I think that this requirement can also cover the kind of explanatory relationship that must obtain between beliefs about necessary propositions and their truth for them to constitute knowledge. See also (Gibbard 2003), Chapter 13 on the notion of "deep vindication" and knowledge in the "more demanding sense" for a related formulation.

inquiry seems to aim at a state where the truth of our normative judgments plays a role in explaining why we hold them.[8]

1.3. Essentially Guiding *and* Descriptive?

A first puzzle about normative judgments in general and ethical judgments in particular is thus that they seem to be descriptive judgments, but they also seem to have a propensity to influence our actions and attitudes that other descriptive judgments lack. How is it that every other kind of descriptive judgment we can think of seems to require a background of beliefs and desires, which we might not have had, to influence our actions, beliefs, and valenced attitudes, but judgments about what to do, think, and feel require no such background? How is it that a propensity to influence our responses seems essential to our normative judgments, but essential to no other descriptive judgments? Since ethical judgments entail that we should have valenced attitudes, and these attitudes involve motivations to do certain things, the puzzle about ethical judgments is often put in terms of their relation to motivation. How can ethical judgments be descriptive if – unlike other descriptive beliefs – these judgments seem to have an essential propensity to motivate us, which does not seem to depend upon our having a contingent desire to be ethical?

One response to this puzzle is that of the anti-descripitivist, who maintains that normative judgments are not, despite certain appearances, descriptive. The anti-descriptivist maintains that the mental states normative claims express are something other than descriptive beliefs that are true or false in virtue of corresponding or failing to correspond to facts that can explain such things as why we hold the beliefs about them we do. Anti-descriptivists are also called 'expressivists' because they think that normative claims are to be analyzed, not in terms of some special descriptive content they possess,

[8] Philosophers like Ayer (1937) and Stevenson (1944) tried to explain all normative inquiry as a kind of applied normative inquiry that took answers to basic normative questions for granted and inquired only into the ordinary descriptive features of circumstances to which these basic answers were to be applied. If this were how normative inquiry operated, it might not presuppose that normative facts can play a role in the explanation of our normative judgments. But as we have seen, we engage in basic normative inquiry all the time, and considerations of the relationship between inquiry, knowledge, and ontic explanation do suggest that this kind of inquiry presupposes that normative facts can play a role in explaining our normative judgments.

but rather in terms of the special, non-descriptive mental states they express. The expressivist explains the unique propensity of normative judgments to influence our responses in terms of their being states that play an action- and attitude-guiding role instead of a representational role. This solves the part of the puzzle about how normative judgments are *un*like other descriptive judgments, but it does not solve the part about how they are like descriptive judgments. To solve our puzzle, the expressivist must give an explanation of what we are doing when we engage in basic normative inquiry that explains away the appearance that we are trying to align our normative judgments with normative facts. Expressivists have had some plausible things to say about why normative debates between people might not actually presuppose normative facts that are capable of explaining non-normative phenomena. But I think they have done little to explain away the appearance that basic normative inquiry presupposes facts about what we should do, think, and feel that are capable of explaining our normative judgments.[9]

Another response to this puzzle is that of the descriptivist about normative judgments. The descriptivist insists that our normative judgments are indeed true or false in virtue of corresponding or failing to correspond to facts about what to do, think, and feel that can – if they obtain – explain such things as why we make the normative judgments we do.[10] Unlike the expressivist, the descriptivist has a natural explanation of why normative inquiry appears to be an attempt to align our normative judgments with normative facts, since according to her, this is exactly what normative inquiry actually is. But the descriptivist might seem to have a more difficult time than the expressivist when it comes to explaining why normative judgments seem to have a propensity to influence our actions and attitudes that other descriptive judgments lack. Would it not, after all, be most simple and explanatory to say that it is no part of the nature of any descriptive

[9] Expressivists have had very little to say about what goes on in basic first person normative inquiry, but we shall examine the accounts suggested by the scant remarks of expressivists like (Gibbard 1990) and (Blackburn 1998) in Chapter 6.

[10] I should point out that while descriptivism so characterized (and the attendant view of what it is for a judgment to be descriptive) is particularly congenial to a correspondence theory of truth itself, it is compatible with the rejection of such a theory. According to a correspondence theory of truth, the truth of a truth-bearer (e.g. thought or sentence) itself consists in the correspondence of the truth-bearer to a fact that makes it true. We could, however, maintain that descriptive judgments are true in virtue of corresponding to facts, but deny that this has anything to do with the nature of truth itself (perhaps it has rather to do with what it is for a state to be a descriptive judgment). A descriptivist could thus happily accept a minimalist or deflationary view of the nature of truth; she simply adds that descriptive truth-bearers are true (in this minimal or deflationary sense) only when they correspond to facts.

judgment to influence attitudes or actions in the absence of contingent beliefs and desires, and to explain the essential propensity of normative judgments to influence our responses as due to their being judgments of a non-descriptive nature?

One thing the descriptivist might do is agree that no descriptive judgments are essentially response-guiding, but deny that normative judgments are essentially response-guiding after all. We might call this position *judgment externalism* about normative judgments, since it denies the *judgment internalist* thesis that it is part of what it is to judge that one should do, think, or feel something that the judgment has a propensity to cause one to do, think, or feel it.[11] Such judgment externalists must explain away the appearance that an essential propensity to guide our responses is part of what sets normative judgments aside as importantly different from merely descriptive judgments. Judgment externalists can admit that our normative judgments do have a special propensity to influence our responses, but insist that this is a contingent fact about our psychology rather than a fact about what it is to make these judgments. Just as it is a contingent fact about the psychology of many mammals that they tend to be made fearful or averse by snake-like images, but no part of what it is to see a snake-like image to tend to respond with fear or aversion, perhaps it is simply a contingent fact about us that we tend to be guided by appearances of what we should do, think, and feel.

Perhaps, but the burden of proof seems to be on the judgment externalist. A special propensity to influence our actions, beliefs, and valenced attitudes seems to be a central part of our grasp on what it is to judge that one should do, think, or feel something. What is it to make a normative judgment, if not in part to be in a state that has this kind of influence? The judgment externalist seems forced to say that normative judgments are what they are simply in virtue of their representing a special kind of fact. We shall discuss difficulties about normative facts below, but there are particular problems with understanding normative judgments solely in terms of the facts they represent. For one thing, there are very good reasons for descriptivists to try to identify

[11] This terminology is derived from (Darwall 1983, 54). I should point out, however, that my characterization of judgment internalism is in an important way weaker than many formulations in the literature. Many such formulations of judgment internalism hold, for instance, that it is part of what it is to judge that one shouldn't do something or that doing it would be wrong that one is *actually motivated* not to do it. My formulation demands no such thing – to judge that one shouldn't do something has a propensity to motivate one not to do it, but as we have seen this propensity might not actually succeed in causing one to have the motivation.

normative facts with facts that we can describe in non-normative terms. Certain kinds of Utilitarians might want to say, for instance, that facts about what we should do are nothing but facts about what would maximize the happiness of all sentient beings. But it certainly seems that you can represent an act as maximizing universal happiness without thereby representing the act as the thing to do, and similar remarks seem to go for representations of any other kind of fact that we can identify in non-normative terms. The most plausible explanation of the difference is that normative judgments have, whereas representations of these other facts lack, an essential propensity to influence our responses.

Fortunately, descriptivists need not be judgment externalists. They can maintain that representing a certain kind of fact is *part* of what it is for a mental state to be a normative judgment, but maintain as well that representations of this kind of fact will not count as normative judgments unless they also play the right kinds of roles in influencing our actions, beliefs, and valenced attitudes.[12] As Frege taught us long ago, distinct thoughts can represent the same things under different guises or modes of presentation. A judgment on Richard Nixon's part that MARK FELT IS A MEMBER OF THE FBI seems distinct from a judgment on his part that DEEP THROAT IS A MEMBER OF THE FBI – Nixon might have confidently made the former long before he came to suspect the latter.[13] But both judgments would have been made true by the same fact – namely that Mark Felt (aka 'Deep Throat') was indeed a member of the FBI. Similarly, Jimmy Carter's making the judgment I WAS PRESIDENT is distinct from his making the judgment JIMMY CARTER WAS PRESIDENT – were Carter to forget his identity and his days in the White House, he might make the latter judgment as a result of reading some history books without making the former judgment (he might even make the judgment I WAS NEVER PRESIDENT given the unlikelihood that he was in light of his evidence).[14] Still, both Carter's pre-amnesic judgment I WAS PRESIDENT and his post-amnesic judgment JIMMY CARTER WAS PRESIDENT would be made true by the same fact – namely that he, Jimmy Carter, was president.[15]

---

[12] See for instance Wedgewood (2004) and Tresan (2006).

[13] My convention throughout will be to use small-caps to denote concepts and thoughts (or judgments), and to use italics to denote properties and propositions (or states of affairs or facts).

[14] See for instance Perry (1979) and Lewis (1979).

[15] On my preferred way of thinking about these things, a fact is a state of affairs or a true Russellian proposition, where the latter kinds of entities are structured complexes of objects and properties (conceived

We can think of pairs of judgments like Nixon's <MARK FELT IS A MEMBER OF THE FBI, DEEP THROAT IS A MEMBER OF THE FBI> and Carter's < I WAS PRESIDENT, JIMMY CARTER WAS PRESIDENT> as different modes of presentation of - or distinct ways of representing - the same fact. What the descriptivist about normative judgments can say, then, is that our judgments about what to do, think, and feel are special modes of presentation of facts about what we should do, think, and feel, which both represent these facts and essentially guide our actions, thoughts, and feelings.[16] The idea of states of

---

of as universals). Thus, a fact is thus neither a true thought nor a set of possible worlds that includes the actual world. To my mind, either of these alternative ways about facts would be disastrous from the standpoint of the role facts play in ontic explanations (and the role our judgments about them play in epistemic explanations). There are no facts about Mark Felt that could fail to play the same causal or explanatory roles as facts about Deep Throat (observe that because thoughts involving DEEP THROAT refer to Mark Felt / Deep Throat rigidly, states of affairs where someone else is deputy director of the FBI, passes information to Bob Woodward, etc. are not states of affairs involving Deep Throat.) But, to take a somewhat contrived example, it seems that the fact that a figure has four sides can explain things that the fact that it has four angles cannot explain – like what makes it the case that it has more sides than a triangle.

[16] There is an important way in which quasi-Realist expressivists can hold something quite similar to this position, so I should take care to explain how this is different and what exactly makes it a descriptivist position. An expressivist could combine the following ingredients: (i) a deflationary view of truth according to which (for instance) the only thing there is to our concept of truth is that to think something true is to agree with it, (ii) the view that facts are just true propositions, (iii) a coarse-grained view of propositions according to which facts are sets of possible worlds in which they are true, and (iv) the view that a mode of presentation of a fact is a mental state that is made true by that fact, but where the truth-making need not be a matter of the mental state's *representing* that fact – the fact could, for instance, make it the case that the thought is true in the sense that the fact that touching a stove will be painful makes it the case that I should not touch it. Following Gibbard (2003), let us call the more expansive sense in which a mode of presentation can be made true by a fact 'the mode of presentation's signifying the fact'. What a quasi-Realist expressivist who accepts (i)-(iv) can say, then, is that "our judgments about what to do, think, and feel are special modes of presentation of facts about what we should do, think, and feel, which both *signify* these facts and essentially guide our actions, thoughts, and feelings" (Were he to accept (ii), I take it this is exactly what would be entailed by Gibbard's (2003) view).

The key difference between this sort of quasi-Realist expressivism and the kind of judgment-internalist descriptivism described in the text is thus whether normative judgments are held to be made true in virtue of representing normative facts or in virtue of signifying such facts in a non-representational way. This corresponds exactly, I think, to the general way in which we should understand the dispute between expressivists and descripitivsts about normative judgments. A descriptive judgment is a judgment that represents facts or states of affairs the obtaining of which make the judgment true and the failure of which to obtain make the judgment false. (Judgments, of course, are not the only kinds of mental states that can represent states of affairs – sensory and perceptual states do so as well. Judgments can be distinguished in they play such "domain general" roles as being "inferable" or more generally able to serve as premises and conclusions of inferences, being states that can be in literal agreement or disagreement with one another, and interacting with "domain general" conative states like desires or valenced attitudes more generally (rather than, say, reflex mechanisms) to produce behavior). What descriptivists about normative judgments affirm, and expressivists about normative judgments deny, is exactly that normative judgments are made true or false in virtue of representing states of affairs that obtain or fail to obtain.

What it is for a truth-bearer to represent its truth-maker (the question of intentionality) is one of the most difficult and fascinating questions studied by philosophers, but the fact that we cannot (now) explain exactly what it is for one state to represent another certainly does not mean that we cannot rely upon the notion of representation. (Indeed, what seems to make the notion of representation so important

mind that both represent a kind of fact and play a role in interacting with other mental states that alternative modes of presentation of those facts do not is by no means new. A relatively popular position in the philosophy of mind holds that features of our subjective experiences (like the red appearance of an apple) represent facts about the objects of experience (like that the apple's being red), but interact with other mental states in ways other representations of the same facts do not.[17] Indeed, quite a few theorists have proposed that affective experiences like pain are states that both represent bodily conditions (like tissue damage) and have a direct propensity to motivate or cause behavior that alters those bodily states.[18]

One challenge for descriptivists who attribute both a representational and a response guiding role to normative judgments is to explain how the two fit together. In the case of states like pain, there is a good evolutionary explanation of why we would have come to have states that both represent bodily damage and cause us to do something about it. Why, however, would we have come to have a single state that both represents normative facts and influences our responses accordingly? Why, moreover, would the response-guiding state be not merely representational but judgmental or belief-like – the kind of state we can come to as a result of domain general inference as opposed to more direct sensitivity to the states of affairs it represents? Closely related to this challenge are the pressures faced by descriptivists of all stripes to move from the position that normative thought represents normative facts to the stronger position that there actually *are* normative facts. This saddles the descriptivist with metaphysical problems about

---

and difficult is exactly that it appears to be both taken for granted by and absolutely central to so very much of our philosophical theorizing.) We could rely upon representation to distinguish descriptive from non-descriptive states of mind even if it turns out to be a primitive or normative relation. It should be mentioned, however, that most leading theories of representation entail that a mental state represents a state of affairs just in case there is a causal or explanatory relation between that kind of mental state (or its components) and that kind of state of affairs (or its components) (See for instance Stampe (1977), Dretske (1981, 1988), and Fodor (1987, 1990)). This could provide a deeper theoretical rationale for distinguishing descriptive from non-descriptive mental states as those that are made true or false by the kind of facts that can figure into the explanation of our beliefs about them.

[17] See for instance Lewis (1980), Lycan (1987), and Tye (1995, 2003). Incidentally, the idea that a qualitative state has functional roles that other states that represent the same facts might lack does not imply "strong representationalism" about qualia; it rather implies only "weak" representatioanlism, which makes no commitments to whether the qualitative states can be reduced to certain functional states with a particular representational content – only that qualitative states such representational content.

[18] See for instance Lycan (1987) and Hall (2008).

what kind of things normative facts are and how they can fit into our best picture of what there is.

The mere fact that a kind of thought or talk is descriptive does not mean that there actually are non-trivial facts about its subject matter. Ancient Greek thought about such deities as Zeus was descriptive, as was 17th and 18th century thought about phlogiston, but there are of course no facts about Zeus and phlogiston other than the fact that these entities do not exist. What this means, of course, is that Zeus-thought and phlogiston-thought was mistaken and untrue. We can thus be descriptivists about a kind of thought without committing ourselves to the existence of the facts that the thought seeks to track if we are prepared to be error theorists about that kind of thought who think that all such thought is mistaken. A descriptivist about normative judgments could thus avoid having to admit that there actually are normative facts about what to do, think, and feel if only she were willing to be an error theorist about normative thought. Interestingly enough, some philosophers have been willing to embrace error theory about certain special domains of normative thought – perhaps most notably ethical thought and thought about which valenced attitudes we should have.[19] But there seems to be little willingness to embrace a general form of error theory about normativity. I know of no one who is an actually error theorist about what to believe or what to do – no one, that is, who genuinely thinks that nothing (including the avoidance of one's own pain) counts in favor of doing anything, or that no states of belief (including suspensions of judgment) are ever any more justified than others.

I believe that the greater willingness of many to be error theorists about ethics and reasons to have valenced attitudes stems from an underappreciation of two things. The first is the way in which ethical thought and thought about what valenced attitudes to have is integrated with practical thought, or thought about what to do. The second is the way in which all normative thought - including thought about what to do and what to believe - has the very same kinds of puzzling features that lead to puzzles about ethical thought and thought about what valenced attitudes to have. One of the things I will be trying to do in this dissertation is to elucidate the connections between ethics, valenced

[19] See for instance Mackie (1977) and Joyce (2001).

attitudes, and reasons for action. I will also be trying to show how the puzzling features shared by all normative judgments require that we give these judgments a similar theoretical treatment. The understanding I will develop of the relationships between ethical thought, valenced attitudes, and normative thought in general will, I believe, show error theory about ethics and reasons for valenced attitudes to be just as hard to swallow as a general error theory about what to do and believe. But for now it will suffice to observe something about why error theory about what we should do and believe is so hard to accept, and to note that error theory about ethics and reasons for feeling is difficult to accept for much the same reasons.

We live our lives in normative terms. We can scarcely make a decision or act without relying on thoughts about what we should do, and we cannot get very far in thinking about what the world is like without thinking about which conclusions our evidence actually supports. Indeed, to arrive at her position the error theorist about reasons for belief would seem to have to suppose that there exist what her own theory denies – namely reasons to believe her theory. Thoughts about certain supernatural beings or certain theoretical entities are thoughts that we can admit to be erroneous and learn to live without. But we don't seem to be able to afford to admit that all thought about what to do and believe is mistaken – if it were, we could not even have non-mistaken thoughts about what to do in light of the mistake.

Of course, the mere fact that it is difficult or even impossible to live without a kind of thought is no evidence that the thought in question is correct. But there are at least three reasons why the central place of normative thought in our lives makes error theory about it a theoretically dissatisfying position. First, because life will be much easier if we can remain confident that some things really are worth doing and some states of belief really are justified, philosophers will be much more interested in seeing if some account of normative judgments that is not error-theoretic can be made to work. Second, if we did become convinced that error theory was true of our current normative thought, we would probably develop a kind of thought that played much the same roles but was not doomed to error, and this slightly reformed kind of thought would become the object of philosophical interest in normative judgments. Finally, we shall see later on that certain descriptivist theories of normative judgment lead to error theory in ways that

appear to show that the theories, rather than our normative judgments, are mistaken. These theories say that our normative judgments are about facts of a certain kind *K*, but the considerations that show there to be no facts of kind *K* do not seem to show that nothing is really worth doing or that no beliefs are any more justified than others. A plausible explanation for this is that a judgment does not have to be a belief about facts of kind *K* for it to play the central roles in our lives that are definitive of normative judgments.

It might seem as though it would be easier to do without ethical thought than it would be to do without all thought about what to do and what to believe. While this is not something I wish to challenge just yet, we can still note two things. The first is that, as we have seen, ethical thought plays a central role in both our everyday decisions and our farther-reaching choices about what to do with our lives. Even if its role is slightly less central than that of normative thought in general, error theory about ethics should be hard to accept for the same reasons that error theory about normativity is hard to accept. Second, if we are not going to be error theorists about reasons for action in general, it can seem rather arbitrary to be error theorists about ethical reasons for action in particular. What is it about ethical reasons that is supposed to make them so much more suspicious than reasons to, say, preserve our own health or avoid our own pain? It is sometimes said that ethical (or, more specifically, moral) reasons for action are "categorical," or reasons for us to do things whether or not we want to do them. But surely this is how most of us tend to think about other reasons for action too – we don't say that a depressed person has no reason to avoid her own pain or preserve her own health just because she doesn't feel inclined to do so. Of course, we might turn out to be wrong about the categorical nature of prudential reasons. But why couldn't we just as easily turn out to be wrong about the categorical nature of ethical reasons?

1.4. Normative Facts?

A descriptivist thus faces a second puzzle about normative judgments in general and ethical judgments in particular. On pain of error theory, the descriptivist must maintain

that there actually are normative and ethical facts, but there are problems about what these facts are and how they fit into our best understanding of the kinds of things that exist. Those who worry about how there could be such things as normative facts are sometimes accused of adhering to a kind of dogmatic naturalism, or assuming without justification that the only things that exist must be "naturalistic" in character. While I must confess that I'm not quite sure what it is for something to be "naturalistic," I think that I understand enough of what is meant by the charge to say that it is entirely mistaken. It makes perfect sense to worry about normative facts even if you are a dualist about the mind or Platonist about mathematical entities who believes in such "non-natural" things as non-physical mental properties and abstract objects. To worry about normative facts we do not have to assume that only one kind of thing exists. We need simply to attend to why we think that certain kinds of things do not exist.

According to a view we might call minimalism about facts, to say that something is a fact is to do little more than assert it.[20] If minimalism is true, and expressions of mental states other than descriptive beliefs can be assertions, then there will be room to talk about non-descriptive facts, or facts of a kind other than those which can be *represented* by states like belief. While I suspect both that minimalism is false and that only expressions of descriptive beliefs can be assertions, a detailed discussion of these issues is beyond the scope of this dissertation. But both those who do and those who do not believe that there can be non-descriptive facts should be able to agree on certain epistemic principles for when we should believe that *descriptive* facts obtain.

In particular, fact-theorists of all stripes should agree that considerations of explanatory parsimony support believing that there exist only those descriptive facts that either figure into our best explanation of the total phenomena or else get entailed by this explanation. Facts that figure directly into our best explanation of the way world is presumably include those discussed by fundamental physics. Facts that are entailed by our best explanation of what there is that are not themselves very important parts of it include facts about the average weight of males living in the U.S. in 1990.[21] All other descriptive facts either pull their own explanatory weight or get entailed by others. Many

[20] See for instance Gibbard (2003, 182).
[21] The example of averages in this regard is due to Harman (1977).

philosophers think that facts about color, heat and chemistry are entailed by physical facts, and those of us who are physicalists about the mind also think that mental facts are entailed by physical facts.[22] Of course, if you are a dualist about the mind you will think that mental facts pull their own weight in explaining our experiences and thus enter directly into our best explanation of what there is without being entailed by anything else.

There must, however, be some constraints on what will for the sake of the parsimony principle be allowed to count as the "total phenomena", or else it would have no teeth. We might think, for instance, that if we can explain such phenomena as thunder and lightning, the misfortunes of dishonest traders and those who were inhospitable, and the emergence of the Balkan and Rhodope mountains without any reference to facts about Zeus, then we have reason to believe that there are no such facts. But what if the defender of Zeus-facts were to object that the *total* phenomena include such phenomena as Zeus turning into a bull and raping women, and we *do* need facts about Zeus to explain *these* phenomena?

A good response seems to be the following. Facts about Zeus are not only unnecessary for explaining such phenomena as thunder and lightning, the misfortunes of the dishonest and inhospitable, and the emergence of certain mountain ranges. They are also unnecessary for explaining why people believed that Zeus turns into a bull and rapes women, and why they believed all of the other things they believed about Zeus. The general lesson seems to be that considerations of parsimony dictate the following. If we do not need a certain kind of descriptive fact to explain anything else, and we can best explain all of our beliefs about such facts without invoking them (or an explanation that entails them), then we should not believe that there are any such facts.[23]

---

[22] For an excellent treatment of the case that all these kinds of facts are entailed by our best explanation of what there is, see (Jackson 1998).

[23] See for instance (Harman 1977) and (Gibbard 1990, 2003). David Enoch (2007) has recently objected to this criterion, arguing that it is enough if belief in a kind of fact is indispensible to a "non-optional" project for it to be the case that we should believe that it exists. Enoch says that by 'non-optional' he is unsure whether he means projects from which "we *can*not disengage, or rather those we should not disengage, or perhaps some combination of the two." I am quite unsure what Enoch means by a project 'we cannot disengage'; whether read as a claim about psychological, metaphysical, or conceptual impossibility it does not seem that either of his two examples – deliberation and explanation – really qualify. I am also quite unsure as to why it would be at all plausible to claim that simply because we should engage in project *P* and project *P* requires belief in facts of kind *F* that we have *epistemic* reason to believe in facts of kind *F*. But what really baffles me is how, given that he is a descriptivist and not an expressivist quasi-realist,

Do we need normative facts to explain anything? We certainly do not seem to need facts about what anyone should do, think, or feel to explain anything about those parts of the world that are outside of the influence of practical, epistemic, and ethical agents – those beings who actually should do, think, and feel things, or who have and tend to be guided by thoughts about how they should respond.[24] Indeed, if we need normative facts to explain anything, it seems that it must be something about the attitudes, behavior, or normative thinking of such agents.[25] Now, as we have seen, agents' *views* about what they should do, think, and feel play an important role in explaining their attitudes and behavior. But do we really need *facts* about how an agent actually should respond in order to explain her responses?

An agent's views about what she should do, think, or feel seem to explain her actions and attitudes equally well whether or not they are actually true. In order to explain why an agent has the views she does about how she should respond, we may need to cite various facts about how that agent was acculturated, including of course the normative views of those around her. But follow the chain of acculturation back, and it looks as though you need explain only how tendencies to make certain normative

---

Enoch can think that the deliberative project – that of figuring out what to do and why – is anything other than a sub-element of the explanatory project of figuring out what is the case and why.

[24] Some people may well think that, say, physical facts have normative explanations – they may think, for instance, that the world is finely tuned because that supports sentient life and sentient life is good. But these people presumably think that the basic structure of the physical world is within the scope of influence of certain agents, for instance the Judeo-Christian deity.

[25] Some people who believe in forces like that referred to as 'karma' might seem to think that facts about who has done, say, right and wrong can explain certain events (like natural disasters befalling certain people) without explaining why anyone (including deities) does, feels, or judges anything. One thing that these people might have in mind is that certain features of acts people perform (e.g. being an act of torturing an innocent for fun) both figure into karmic laws and make those acts wrong, in which case this would not be an instance of thinking that the explanatory role is played by the normative facts themselves (for an account of this kind of "normative explanation" see for instance (Gibbard, 2003, Chapter 10)). Alternatively, advocates of these kinds of karmic explanations might think that normative facts are reducible other kinds of facts (say facts about what causes suffering), and that these facts figure into the karmic laws. We will discuss reductive accounts of normative facts shortly, and the points made there will apply to these kinds of views. Another possibility is that some people think that certain irreducible features of actions figure into karmic laws, and describe these features with ethical language, but do not actually think that these are normative features, or features the instantiation of which entails (i.e. guarantees in virtue of meaning and logic alone) something about what we should do, think or feel. What seems difficult to make sense of is the view that irreducible facts about what to do, think, or feel figure into karma-like explanations in ways that do not explain anyone's responses.

judgments got passed down by mechanisms of biological or cultural evolution.[26] In the case of our capacities to form beliefs about such things as tables, chairs, and electrons, we need to posit the reliability of these mechanisms in tracking their subject matter to explain why they would enhance survival and reproduction and thus get passed down. Do we need to posit the reliability of our capacities to make normative judgments in tracking normative facts order to explain why these capacities were adaptive and how they got passed down?

The answer might seem to depend upon whether we think that normative facts are *reducible* to some other kind of facts. We tend to think, for instance, that facts about water turned out to be identical to facts about H_20. In the same way one might think that normative facts about how we should respond are identical to facts about our responses that we could pick out in non-normative terms. An egoistic hedonist might think, for instance, that the fact that you should perform an act is nothing but the fact that the act will promote your own pleasure.[27] On the other hand, it is natural to think that certain facts, like those about the most fundamental physical particles in the universe, are irreducible to any other kinds of facts. One might think that normative facts are irreducible in much the same way.[28]

---

[26] I doubt that anything really hangs on the details of the true story of how we came to make the normative judgments we do. The same issues I shall raise about normative facts could be raised if we had been set up to make these judgments by deities, or had we been spontaneously generated a few moments ago by lighting hitting a swamp, or what have you I stick to the actual evolutionary story for heuristic purposes.

[27] If you don't like egoistic hedonism (as I don't), you can easily substitute your favorite theory of what we have reason to do in its place. Thus we could consider the view that facts about what we should do are identical to facts about what would promote universal happiness, or score highest on a certain parameterization of Ross's list of "prima facie" duties, or what have you.

[28] I should clarify that a higher level fact's being reducible to a lower level fact in this sense is a stronger criterion than the higher level fact's simply being entailed by the lower level fact. The reducibility of a higher level fact to a lower level fact requires that the two facts be identical. A lower level fact could, however, entail a higher level fact without being identical to it if there is more than one way for the higher level fact to be realized by lower level facts. This is what many of us think about neural facts and mental facts: that the facts about the way our brains are entail facts about the way our minds are, but that they are not identical to such facts, since we could have been made of silicon and had silicon states rather than neural states and still have had the same mental states we do.

If the normative facts were entailed by but not identical to facts that we could specify in non-normative terms, then the normative facts could still play a role in our best explanation of what there is that is not exhausted by the explanatory contributions of the non-normatively specified facts. But the fact that we only need facts that are specified in non-normative terms in order to explain our normative judgments means that normative facts will be explanatorily superfluous unless they are *identical* to some non-normatively specified facts.

If normative facts are irreducible, then we pretty clearly do not need them in order to explain the evolutionary origins of our normative capacities. In order to explain how agents' capacities to form normative judgments enhanced their survival and reproduction, we need only explain how these capacities caused agents to do, think, and feel certain kinds of things that it was adaptive for them to do, think, and feel in ancestral environments. These kinds of things would presumably include actions that would have tended to prevent bodily damage, attract mates, and position offspring to survive and reproduce, beliefs that would have been accurate enough on balance, and emotions that would have facilitated social cooperation. But the irreducible fact (if it was a fact) that these were the kinds of things that agents actually had reason to do, think, or feel is completely irrelevant to the explanation of why they were adaptive. We simply do not need the fact that an agent should respond in some way in addition to the fact that the response would have such-and-such tendencies to prevent bodily damage, attract a mate, etc. in order to explain why tendencies to judge that one should respond in that way were adaptive.[29]

It looks, then, as though we will need to posit normative facts that were successfully "tracked" in evolution only if such facts are reducible to other facts that evolution did design our normative judgments to track. Return to our egoistic hedonist, who thinks that the fact that you should do something is identical to the fact that your doing it will promote your own pleasure. If she is right, and evolution designed our normative capacities to track facts about what will promote our pleasure,[30] then it follows that evolution designed our normative capacities to track facts about what we should do. In this way the reductivist about normative facts can identify them with facts that we already need to explain our normative judgments. If the fact-identities can be sustained, this seems to give us a promising way to explain how normative facts figure into the

[29] For arguments along these lines see for instance Blackburn (1988), Gibbard (1990), and Street (2006).
[30] This might seem pretty incredible, but suppose that by the time capacities to make and be guided by judgments about reasons for action came on the scene, the best the mechanism could do to enhance fitness was to take advantage of existing adaptive correlations between hedonic states and reproductive success. The point is not so much the plausibility as the form of the reductivist's explanation. We could equally consider a crazier reductive view (e.g. that facts about what we should do are facts about which acts will maximize the number of our descendants) together with a more plausible evolutionary story to make the same points.

explanation of our normative judgments. But how can we tell if a given reductivist proposal is correct?

There are two general ways in which normative facts could be reducible to facts like those about what will promote one's own pleasure. First, the identity between facts about what to do and facts about pleasure could be a matter of meaning and logic alone, like the identity of facts about brothers and facts about male siblings. In this case facts about what to do would be *analytically reducible* to facts about what would promote one's pleasure, or identical to such facts because judgments about what one should do can be analyzed as a kind of judgment about what would promote one's pleasure. A second way in which normative facts could be identical to facts like those about pleasure would be for the identity between these facts to be a synthetic truth that is not simply a matter of meaning and logic. This is how most of us think about the identity between facts about water and facts about H_20 – while judgments about water (conceived of as WATER) pick out the same facts as judgments about H_2O (conceived of as H_2O), this is no part of the meaning of these judgments. In this case facts about what one should do would be said to be *synthetically reducible* to facts about what would promote one's pleasure.

The problem with claiming that normative facts are synthetically reducible to other kinds of facts is that synthetic identities between normative facts and other facts are explanatorily superfluous. As with other kinds of descriptive facts, we should only believe in facts about identities if they enter into (or are entailed by) the best explanation of something - at the very least that of our believing in such identities. We should, for instance, believe that facts about water are identical to facts about H_20 because this enters into (or is entailed by[31]) our best explanation of what water is like and how we came to have the beliefs about it we do. But we should not, for instance, believe that facts about Mt. Olympus are identical to facts about Zeus's actual home because no such identity enters into (or get entailed by) our best explanation of how things are. An identity between normative facts other facts, like those about our pleasure, would add nothing to our evolutionary story or any other part of our explanation of why people make the

[31] See (Lewis 1970) and especially (Jackson 1998) for a powerful case that this identity follows by analytic entailment from our best theory.

normative judgments they do. To explain how it was adaptive to judge that you should pursue pleasure, we would not need an identity between the fact that something was pleasurable and the fact that you should do it over and above facts about how pursuing pleasure in ancestral environments would have promoted health, led to reproductive encounters, etc.

It is actually quite popular for descriptivists about normative judgments to eschew conceptual analysis and insist that normative judgments cannot be analyzed in any further, non-normative terms. For these *non-analytic* descriptivists, normative facts must be either irreducible or synthetically reducible to other kinds of facts. But we have seen that we do not seem to need either irreducible or synthetically reducible normative facts to explain how we came to make the normative judgments we do. If this is right, then considerations of parsimony would appear to dictate that non-analytic descriptivists are committed to error theory, or to saying that all of our judgments about what to do, think, and feel are just as mistaken and untrue as beliefs about Zeus.

While we have already seen how unappealing this sort of error theory is, it is worth nothing how this result seems to suggest that non-analytic descriptivism has got things wrong about our normative judgments. The fact that we do not need to posit a Zeus to account for judgments about Zeus seems to be a perfectly good reason to embrace error theory about Zeus. Intuitively, this seems to successfully explain how Zeus-judgments have got things wrong, and we do not feel somehow tempted to say instead that something has gone wrong with our account of Zeus-judgments. But the mere fact that we do not need a special kind of fact or fact identity to explain our judgments about what to do or believe does not seem to mean that nothing is really worth doing or that none of our belief states are ever more justified than others. We have seen that reflective equilibrium methods of normative inquiry rely heavily on debunking considerations –explaining away an intuition by showing it to be generated by processes that do not reflect the normative facts. But the general irrelevance of special facts or identities to the explanation of all normative cognition does not seem to hand us a genuine debunking explanation of all of the intuitions that we shall ever consult in basic inquiry about what to do or believe. This should sound a warning bell for the non-analytic descriptivist. For even if she thinks, *per impossible*, that we actually need more

facts or identities than we have dreamed of in our explanations of normative judgments, the apparent irrelevance of whether this is so to whether our normative cognition gets things right should strongly suggest that non-analytic descriptivism gets things wrong about our normative judgments.

If non-analytic descriptivism about normative judgments leads to an unacceptable form of error theory, then the only other option for descriptivists is to embrace conceptual analysis. This kind of *analytic descriptivism* maintains that it is a fact about the content of our normative judgments, or the kinds of mental states they are, that they represent a kind of fact that we can describe in non-normative terms. If the analytic descriptivist is right about the content of normative judgments, then normative facts analytically reduce, or are identical in virtue of meaning, to these facts that our normative judgments represent (assuming these facts exist).

Analytic descriptivism about normative judgments, and ethical judgments in particular, has been out of favor throughout much of the last century. While there is no doubt that this has been due in part to an accurate appreciation of the problems that confront the approach, there are reasons to suspect that these problems were sometimes exaggerated. G.E. Moore's (1903) original presentation of his famous "open question argument" against analytic descriptivism seemed to point out little more than that it is not obvious that any descriptivist analysis of ethical judgments is correct. Similarly, Moore's accusation that this kind of descriptivism had committed a "naturalistic fallacy" seemed on reflection to accuse the view of nothing more than thinking that conceptual analysis (or the analysis of ethical concepts in particular) was possible.[32]

The possibility of conceptual analysis is, of course, a highly controversial topic among philosophers, with doubts about the possibility tracing most prominently to Quine (1951). The issues here are complex, but a few observations are of particular importance. First, conceptual analysis of the kind in which we are interested need not take the form of giving synonyms, or linguistic expressions that have exactly the same meaning as those that are usually employed to express the concept under analysis. Conceptual analysis in the sense that is important to us requires only that we can say something informative

[32] See for instance Frankena (1939) and Nakhnikian (1963).

about what it is to make judgments that involve the concepts under analysis. Expressivists are engaged in the analysis of normative concepts in this sense when they explain normative judgments as non-representational states that have certain propensities to influence our responses. The expressivist might well think that there are no expressions in English other than canonical normative expressions (like 'one should do *X*', 'the belief that *P* is justified', 'there is reason to feel ashamed', etc.) that express these concepts, or do so in a way that sheds any more light on our normative concepts than our usual normative language.

Similarly, the descriptivist might think that there are no interesting alternative linguistic expressions that express our normative concepts, but think that we can say things about these concepts for all that. As we have seen, the descriptivist might think that normative judgments are states that represent normative facts under a mode of presentation that has a propensity to influence our attitudes accordingly. She might well think that no other expressions in English have this kind of response-guiding mode of presentation of the normative facts. But this does not prevent her from saying that it is part of what it is for a mental state to be a normative judgment that it represents facts that we can describe in non-normative terms.

By way of analogy, suppose that we had a concept, SCHNIXEN, which had the same descriptive content as VIXEN but represented that content under a "cuddle-guiding" mode of presentation. To judge that someone is a schnixen is to be in a state that represents the fact that she is a vixen, but that also motivates one to cuddle her – a state does not count as a schnixen judgment unless it plays both this representational and this cuddle-guiding role. We don't actually seem to be set up to make schnixen judgments; much as we might be inclined to cuddle vixens, there does not seem to be a single state in us that both represents vixenhood and motivates cuddling.[33] But suppose that we could, and that 'schnixen' was our conventional way expressing SCHNIXEN. Suppose further

[33] Our cuddling motivations seem to be sensitive to such cute-making features as big eyes, big ears, a bushy tail and a whiskery pointy face, but they do not seem to be sensitive to having a particular biological essence. We would seem to be just as inclined to cuddle beings of a different species or sex who were psychologically and morphologically identical to vixens, and this does not seem to be a result of misrepresenting them as vixens. (Moreover, it might not be part of what it is to be in the states in us that represent cute-making features to be motivated to cuddle the beings who we take to have them. Perhaps cute-making features are represented by us under the same modes of presentation as they are represented by sociopaths who have no motivations to cuddle the beings who are represented as having them.)

that the concepts we expressed with 'vixen' and 'female fox' continued to have only their representational roles without the additional cuddle-guiding role, so that our making the judgments we would express as 'Smanatha is a vixen' or 'Samantha is a female fox' had no necessary effect on our motives, but our making the judgment we would express as 'Samantha is a schnixen' motivated us to cuddle Samantha. It seems, then, that what we would express when we said 'A being is a schnixen if and only if she is a female fox' would be an analytic truth - it would be a fact about the kind of state schnixen judgments are that they can be truly made (by those who make them) of all and only those beings who are female foxes. This would be so even though 'female fox' would be no synonym for 'schnixen' – the latter would represent female foxes under a cuddle-guiding mode of presentation that the former does not, and this difference in mode of presentation constitutes a difference in meaning.

A second thing to bear in mind is that the analytic descriptivist's project need not be to capture every exact detail of the normative concepts we actually employ. The project of describing our actual concepts may shade almost imperceptibly into the project of describing concepts that play the most central and interesting roles of our concepts but do not quite match them. Quine (1960) himself seemed to look favorably upon a kind of project much like the latter, which he described as giving a 'paraphrase' of a concept of interest. There are, however, two important caveats about the extent to which analyses of our most basic normative concepts can be understood as paraphrases or proposals for conceptual reform. The first is that most reforming proposals about which concepts to employ are themselves cashed in terms of our basic normative concepts of what to do – they are proposals about what kinds of things we should bother to think about, and what kinds of inquiries we should bother to undertake. If the very notion of WHAT TO DO is at issue, then there is no neutral normative notion in which to cast a proposal for reform. Perhaps the gravitation to the victorious analysis will have to be non-rational – its ability to capture certain features of the pre-theoretic notion will be pointed out, and we simply will (or will not) fall into thinking in terms of it. Alternatively, we might hope that a single proposal for reform could actually be endorsed on the analyses of normative concepts provided by its rivals – that is, that we could say things like 'we should think about reasons to do things in terms of analysis *V* for reasons *X*, *Y*, and *Z*', where this

would come out true on any of the leading proposals for how we should understand talk about what we should do.

The second caveat is that in thinking about reforming analyses of normative concepts, we need to think about what (if anything) should distinguish analyses of normative concepts from substantive theories of what falls under these concepts. There are, I believe, two quite distinct (though related) explanatory projects. The first is that the task of *normative theory* construction, which seeks to answer the basic normative questions "what should we do, think, or feel in circumstances of a given kind?" These take as data our normative intuitions – spontaneous appearances about what to do, think, and feel - at all levels of generality, ranging from intuitions about how to respond in particular cases to intuitions about the plausibility of principles about how we should respond in general. Normative theories sift through which of our intuitions are veridical and which are illusory, and seek to explain more specific normative facts in terms of more general normative facts. The normative regularities we discover might or might not be very straightforward and might or might not be very informative. But even particularists about normative reasons will think that there is something to explain about what there is reason to do, think, and feel. They will presumably think we can explain local normative regularities in terms of statements about what kinds of considerations are generally or *ceteris paribus* reasons to respond in various ways, and which are generally considerations that defeat, weaken, or strengthen certain other reasons.[34] At the very least, she will want to explain the way in which relationships between reasons resist characterization by hard and fast rules – for instance because there are quite generally various kinds of considerations that defeat (or weaken) the case for doing something without counting in favor of doing the opposite.

But our normative theorizing naturally gives rise to questions about *what we are doing* when we are engaged in it, and what the normative theories we come up with are theories about. We want to know *what makes* an answer to our questions about what to do, think, and feel *true*, and how the intuitions we rely upon in answering such questions can provide us with normative knowledge. Questions of this kind about ethical thought in particular are standardly called *metaethical*, and we might call these kinds of questions

[34] See for instance Dancy (2004).

about our normative theorizing more generally metanormative. The task of *metanormative theory construction* is thus that of explaining what normative thinking is up to, what normative truth or correctness consists in, and how methods of normative inquiry can succeed in getting things right or wrong.[35] Metanormative theory construction relies upon a different kind of evidence than that we use to (directly) answer our basic normative questions. Rather than intuitions about what to do, think, and feel, metanormative theories take as data our intuitions about *what it is to make a judgment* about what to do, think, or feel. These include the intuitions that we relied upon earlier which suggested that it is in the nature of judgments about what to do, think, and feel that they have a propensity to influence our actions, beliefs, and valenced attitudes. Perhaps less obviously, our intuitions about the explanatory superfluity of irreducible normative facts were also intuitions about the peculiar content of normative judgments. For the

[35] I apologize if the last two elements of this formulation of the enterprise sound biased in favor of a descriptivist (or indeed *reductive* descriptivist) view of normative judgments. These formulations of metanormative questions are intended to be maximally intuitive and to stay as close to possible to the way they arise in common sense thinking (and I fear that the presuppositions of common sense ways of asking metanormative questions – though perhaps not common sense views about the answers to these questions – *are* biased in favor of reductive descriptivism). The expressivist may have to deny some presuppositions of the second and third questions as I have given them, but she offers an answer to the kind of thing the questions are trying to get at. Her basic answer to the first question will (or in any event should) be that normative correctness or truth is *sui-generis* – the only thing to say in general is that the truth or correctness of the judgment that we should do, think, or feel *X* consists in its being the case that we should do, think, or feel *X*. Her basic answer to the second question could be more informative – it could, for instance, be that someone's knowing that we should do *X* consists not only in its being the case that we should do *X* but also in its being the case that we should rely upon that person's views in figuring out whether or not we should do *X* (as suggested by Gibbard (2003)).

These are probably not the kinds of answers that the person asking about the nature of normative truth and knowledge wanted to get. But the expressivist can go on to explain that these are the only kinds of answers we can have because, contrary to what the questioner was assuming, normative judgments do not represent normative facts, and normative facts cannot enter into explanatory relationships with normative judgments. The non-reductive descriptivist might want to say the same un-informative things that the expressivist says, but I believe that her representationalist commitments will not allow her to do so. If we are to believe in normative facts of the kind we represent, we need a substantive metaphysical and epistemology story about what they explain and how they interact with our normative judgments. If the non-reductive descriptivist wants to retain a metaphysical commitment to *sui-generis* normative facts without the additional explanatory baggage, my best advice to her is to convert to quasi-realist expressivism: become an expressivist who is also a minimalist about truth and facthood. (But be warned: if you want your normative facts really *sui-generis*, you cannot accept ingredients (i)-(iv) that I discussed in note 16. I recommend that you take facts to be true Russellian propositions (rather than sets of possible worlds that include the actual world), and properties to be universals that are individuated more finely than ordered pairs of worlds and extensions. That way, when you make normative claims that express your non-descriptive mental states, you signify instantiations of genuinely *sui-generis* normative properties (rather than instantiations of the boring-old non-normative properties that make it the case that the sui-generis normative properties are instantiated. This does, however, require a broader understanding of the signification relation than that discussed in note 16)).

argument did not depend upon which normative views we held – whatever we thought we should do, the irreducible fact that we should do it seemed unneeded to explain our thinking that we should. Explanations of our responses in terms of irreducible normative facts do not simply threaten to be bad explanations; they actually threaten to be incoherent.[36]

Because of the important difference between the projects of normative and metanormative theory construction, the descriptivist faced with the prospect of giving a reforming analysis of our normative concepts should resist the temptation to replace our notions of what to do, think, and feel with her best guesses as to what we actually should do, think, or feel.[37] Suppose, for instance that someone were to emerge from the enterprise of normative theory construction with a complete view about what we should do in general – say the Utilitarian view that what we should do is always whatever will maximize universal happiness. There still seem to be important questions she can ask. What exactly is her Utilitarian theory is a theory about? What makes it the case that maximizing universal happiness is always the thing to do? How did the various (by her lights veridical) intuitions that she relied upon to arrive at her Utilitarian conclusion provide her with knowledge about what to do?

How, moreover, should the Utilitarian understand disputes between herself and those who have emerged from the task of normative theory construction with rival views – say the egoistic hedonist who thinks that what one should do is always whatever will maximize one's own happiness? To settle disputes over the analysis of a concept, we look to intuitions about the coherent uses of the concept, and to settle disputes between rival proposals for reform we look to considerations of which proposal it would be worth

[36] It might be thought that surely we can make sense of the following: Mike went to the store because going to the store was the thing for him to do. But this is not enough to show that we can make sense of it as an explanation that depends upon an unreduced normative fact. Gibbard (2003) has argued plausibly that such explanations are usually short-hand for something like the following: (i) there is a kind *K* of thing that one should do, (ii) if something is of kind *K*, Mike will think that he should do it, (iii) if Mike thinks he should do something he will do it, (iv) Mike's going to the store was a thing of kind *K*, so (v) Mike will go to the store. But the only premises that actually play a role in this kind of (DN) explanation are premises (ii)-(iv); the normative premise (i) is a superfluous add on from the point of view of the explanation of Mike's behavior. To show that the explanation of Mike's going to the store in terms of the irreducible fact that he had reason to go to the store, we would need to show that it is coherent for facts about what Mike should do to play a role in explaining his behavior above and beyond facts about his views about what he should do. But this sort of thing threatens to be unintelligible.

[37] Along these lines I have misgivings about the approach to reform analysis advocated by Railton (1986), Brandt (1979 – or his approach to questions about *right*, anyway) and Jackson (1998).

our while to think in terms of. But surely neither the Utilitarian nor the egoistic hedonist actually have intuitions to the effect that the other party's views are incoherent. And the question of whether we should think in terms of what will maximize universal happiness or our own happiness (or neither) when we think about what to do is just a question of normative theory construction. If we want to understand the question that Utilitarianism and egoistic hedonism were both trying to answer, we cannot replace the question with either 'what maximizes universal happiness?' or 'what maximizes one's own happiness?'.

Metanormative questions should be of more than academic interest to rival normative theorists like the egoist and the Utilitarian, and they should be of even keener interest to the vast majority of us who have nothing like a systematic answer to basic questions about what to do every given circumstance. If someone relies on intuitions that we think are illusory, or we rely on intuitions that someone else thinks are illusory, we should very much like to know what it is for one of us to be right and the other wrong. To remain confident in the veracity of our intuitions and the falsehood of our rival's, intellectual integrity seems to demand that we be able to tell some story about why we are right and why they are wrong. It would seem to be a big help in telling this kind of story to have some handle on what it is to get things right or wrong about normative matters. Moreover, knowledge of what normative truth consists in might in principle help us figure out which intuitions are illusions and which are veridical - less ambitiously, it might at least protect us from being misled in our normative inquiries by irrelevant considerations.

Thus, in giving an analysis of our normative concepts, reforming or otherwise, the analytic descriptivist should respect the standards of evidence of metanormative theory. Her analysis will be plausible to the extent that it makes sense of our intuitions about what it is to make a normative judgment, including intuitions about which normative judgments are coherent. Perhaps the most charitable way to interpret Moore's open question argument is as a kind of test that can be applied to a descriptivist analysis of our normative judgments. A descriptivist analysis will say that part of what it is to judge that one should do, think, or feel something is to represent the fact that the action, belief, of feeling as having such and so features, which the descriptivist specifies in non-normative

terms. The Moorean test checks if it looks coherent to judge that an action, belief or feeling has those features but to deny that it is something one should do, think, or feel.[38] These kinds of intuitions about coherence of normative judgments are part of the primary dataset for metanormative theorizing which the analytic descriptivist should be out to explain. It is thus a mark against her analysis if it fails to capture some of these intuitions, and if she is to remain competitive, she needs to show how the benefits of her analysis are worth these costs. Ideally, the descriptivist should be able to explain away any intuitions that go against her analysis by telling a story about how we could be misled into thinking that the normative judgments in question are coherent when they are not in fact. But whatever story the analytic descriptivist tells to explain away the intuitions will compete with a rival story – one that seeks to show how the intuitions about coherence point to a genuine and centrally important feature of normative judgments that the descriptivist's analysis fails to capture.

The problem for descriptivist analyses is that most of those that we can come up with seem to fare quite poorly by these standards. We have already seen this in the case of the toy reform analyses of judgments about what to do as judgments about what will maximize universal happiness or one's own happiness. It seems entirely coherent (and in many cases really quite plausible) to think that one should do something even though it won't maximize universal happiness or won't maximize one's own happiness. In fact, the analyses seem unable to make sense of any of our basic normative thought and inquiry – including how the intuitions in favor of Utilitarianism and egoism are supposed to be evidence for these positions and how they are supposed to win out over intuitions to the contrary. In the face of these staggering costs, the analyses seem to have little if anything to recommend them.

[38] See Gibbard (2003) for this kind of interpretation of Moore's argument. Of course, as with any conceptual analyst, the set of features the analytic descriptivist points to could be somewhat vague (see e.g. (Chalmers 1996) and (Chalmers and Jackson 2001). For instance, it seems to be a conceptual truth that a person with a full head of hair is not bald, and that someone with no hair is bald, but there is presumably no way in principle to give necessary and sufficient conditions that say exactly how much hair someone has to be missing to count as bald. At a minimum, what the analytic descriptivist needs to give us is a set of purportedly analytic "supervenience conditionals" of the form 'if an act has features *F*, then the act is the thing to do' (where *F* is specified in non-normative terms). The Moorean test would then be applied to the supervenience conditionals, by seeing whether it looks coherent to deny them.

Perhaps some of the most plausible descriptivist analyses would hold, for instance, that to judge that one should do something representing the action as one that one would be motivated to perform if one were under some particular, non-normatively specified conditions. Candidates for such conditions might include those of possessing full information and attending equally and vividly to all of the relevant non-normative facts,[39] or having undergone the kind of "cognitive psychotherapy" discussed by Brandt (1979). Another kind of descriptivist analysis might hold that to judge that one should do something involves representing the action as prescribed by a certain set of abstract rules that we can identify by means of their content. This kind of analysis would liken an act's status as the thing to do to an act's status as permitted by the rules of soccer, where the content of the rules is part of the concept THE RULES OF SOCCER. In the normative case, the rules might be only *prima-facie* directives (e.g. "don't lie", "don't steal", "don't cause pain", etc.), and the thing to do is whatever is prescribed by the "best systematization" of them – say that which gives us a certain tradeoff between simplicity and match to the prescriptions of each directive.[40]

Unlike the foregoing toy analyses, these analyses really do have something to recommend them. The first kind of analysis can portray basic normative inquiry as an attempt to determine how we would respond under the relevant conditions,[41] and the second kind of analysis can portray this kind of inquiry as the search for the best systematization of the analytically given *prima-facie* directives. But as Moorean tests reveal, these kinds of analyses do run up against our intuitions about coherence. It seems that for any non-normatively specified conditions we could come up with, one could coherently judge that one would be motivated to perform an act if one were under those conditions, yet judge that one should not perform the act.[42] Similarly, it seems that for any set of *prima-facie* directives and systematization scheme we could come up with, one could coherently judge that an act would be prescribed by the resulting best systematization, yet judge that one should not perform it. For any given directive it

[39] Which would resemble the kinds of conditions that Firth (1952) uses to characterize an "ideal observer" in the case of moral judgments in particular.
[40] Perhaps this is the best way to understand Jackson's (1998) proposal. For a model of the kind of systematization in question, one might look to Lewis's (1973) best systems account of the laws of nature.
[41] This is the model of normative inquiry suggested by Lewis (1989).
[42] For criticisms of this kind of Brandt's (1979), see for instance Velleman (1988), Gibbard (1990, p.19-21), and Rosati (2000).

seems we can question whether it should carry any weight at all, and the weight of some might seem as though they should yield entirely in favor of the weight of others.

The problem is not simply that these analyses conflict with some of our intuitions about coherence. The problem is that the best explanation of the conflict seems to be that questions about how we would respond under certain conditions and questions about what a given set of rules prescribes are insufficiently ultimate. Whatever commitment we might have to doing what we would under the listed conditions or following the listed rules seems to depend upon something else. We seem to have independent ways to assess whether we should follow these rules or do as we would under these conditions. If this is right, then what we need to be analyzing are the more fundamental procedures by which we assess whether to follow the rules or do as we would under the conditions. The challenge for any descriptivist analysis of our normative judgments is thus to explain how it captures the content of the most ultimate questions we can ask when we deliberate about what to do, think, or feel.

1.5. Why Be Ethical?

Ethical judgments, I suggested, are normative because they entail judgments that we should have certain valenced attitudes like emotions and desires. Because of this we can understand problems about the apparently descriptive yet motivating role of ethical judgments and the problematic status of ethical facts as part of the more general problems about normative judgments that we have surveyed. But there remains an additional problem about the normativity of ethics. As we saw at the outset, ethical facts seem to bear not only on what we should feel, but what we should do. We usually take it for granted, for instance, that we have reason to avoid doing what is wrong and that good outcomes are outcomes we should bring about. To many of us, these assumptions seem like truisms, or claims it would actually be incoherent to deny. But if it is a conceptual truth that we have reason to be ethical, what is it about our ethical concepts and our concept of what we have reason to do that make it so? Surely many of our reasons to do things are not themselves ethical reasons to do them – the kind of reasons we typically

have to avoid our own pain, do what we find enjoyable, and take care of our health and finances are very often like this. The notion of what we should do thus seems to be quite distinct from the notion of what would be morally permissible or apt to bring about an good overall outcome. But then what guarantees that we should avoid doing what's wrong and that we should bring about the good?

Some people have argued that there is actually no guarantee that we have reason to be moral in the absence of its happening to be the case that our wrongdoing will aversely affect our own interests – say because of the sanctions of society, a deity, or the pangs of our own conscience.[43] More generally, it has been held that there is no conceptual guarantee that we have reason to be moral, and that whether there are such reasons depends upon the actual content of morality. Thus, some have suggested that because we have reason to make the world a better place, we will have reason to be moral if morality turns out to be consequentialist and enjoins us to do just that - but that should morality turn out to prescribe behavior that is not optimific, we may well have no reason to be moral. Others might be inclined to think that one has reason not to be the kind of hypocrite who interferes with others' projects but expects others not to interfere with his. It would follow from that that if being moral is a matter of not interfering with others, anyone who justifiably expects others not to interfere with him has reason to be moral. But should morality turn out to require something other than non-interference (or whatever else we currently expect others to do for us), these people might concede that we lack reason to do as morality requires.[44]

Getting things straight about why exactly we have reason to be moral thus seems to matter a great deal to our substantive thinking about morality. Advocates of substantive moral views sometimes combine our strong intuitions that we have reason to be moral with particular explanations of why we do in order to lobby for their positions. Thus, if our reasons to be moral depend upon prior reasons to avoid sanctions, perhaps the best explanation of why we have reason to be moral is that morality requires doing whatever your society or God will punish you for failing to do. There are, of course,

[43] Views of this kind may be traced to Plato's character Thrasymachus (Plato ca.380 B.C.E.) and Hobbes (1651).

[44] Something very much like these kinds of thoughts about hypocrisy is what I suspect continues to attract people to things like Kantianism and Contractarianism understood as substantive normative theories.

many intuitions that suggest that things like harming and enslaving others would be wrong whether or not they were sanctioned by societies or deities. One explanation of why it would be wrong to harm or enslave people even if one will not be punished for doing so is that these things would lead to misery in the world, and morality requires us to bring about the greatest balance of happiness over misery. An alternative explanation might be that harming and enslaving others interferes with their projects, and morality enjoins non-interference. One way in which these kinds of views can try to compete is in terms of their ability to explain our intuition that we have reason to be moral. The Utilitarian view might try to explain these as reasons to bring about everyone's happiness out of a kind of impartial benevolence. The non-interference view might try to explain our moral reasons as reasons to avoid hypocrisy given that we expect others not to interfere with us. If either impartial benevolence or hypocrisy avoidance seems more plausible as the source of our reasons to be moral, it would seem to put one of these views at an advantage.

If, on the other hand, it is actually a conceptual truth that we have reason to be moral, then our reasons to do so will not depend upon the substantive content of morality. We would have *intrinsic* reasons to be moral, of the kind discussed by Prichard (1912) and Falk (1948), which obtain whether or not moral actions benefit us, make the world a better place, or save us from hypocrisy. If we could show how an act's status as morally wrong analytically entails the existence of reasons not to perform it, we would put substantive moral theories out of the business of showing why we have reason to be moral. If we really have intrinsic reason to be moral, we will have reason to do what is moral whatever the true theory of morality turns out to be. This would mean that we cannot advocate for a particular substantive theory of what morality requires on the basis of its ability to explain why we should follow the dictates of morality. I think that this would have important implications for the ways in which substantive moral positions are defended, and indeed for the strength of the case being put forward in favor of certain of these views.

A particularly puzzling feature of the relationship between morality and reasons for action is that an act's moral wrongness seems to entail the existence of *conclusive* reasons not to perform it. If it would be wrong for you to do something, it doesn't just

seem that there is some reason not to do it – it seems that the reasons against doing it decisively outweigh whatever reasons there might be to do it. That we have conclusive reason not to do what is morally wrong seems not only to be a truth; it actually seems to be a conceptual truth. There seem to be genuine problems with the coherence of thinking that it would be morally wrong for you to do *X* but that you should do *X* all the same. But what could it be about morality or reasons for action that would guarantee this kind of conceptual connection? The notion of moral wrongness is quite distinct from the notion of something that one has most reason not to do – one can have most reason not to waste one's time playing video games in cases in which it would be morally permissible to do so. There also seems to be no conceptual guarantee that moral considerations are always overriding. It looks coherent (and actually quite plausible) to think that there could be circumstances in which it would be morally good to sacrifice one's life to spare a stranger a somewhat more painful death, but in which one would lack conclusive reason to undertake the sacrifice. At the same time, the concept of moral wrongness looks distinct form the concept of a moral consideration that happens to outweigh all of one's other reasons. It looks coherent (and perhaps even plausible) to think that someone could have moral reasons to sacrifice his life for others, which (because he has only a blasé life to look forward to) outweigh his reasons not to sacrifice his life, without its being morally wrong for him to fail to sacrifice his life.

1.6. An Overview of What is to Come

In the next four chapters of this dissertation, I present a theory of the relationship between ethics, what we have reason to feel, and what we have reason to do. I begin by defending the view that ethical judgments are nothing other than judgments about our reasons to have valenced attitudes - they entail judgments that we should have pro- and con-attitudes towards certain things because they are *identical* to these judgments about what to feel. Thus, the concept of a good state of affairs is nothing other than the concept of a state that we have reason to desire, the concept of a morally wrongful act is nothing other than the concept of an act that we should feel obligated not to perform, and so on. Some

have contended that this order of explanation cannot be correct on the grounds that ethical concepts themselves are needed to make sense of judgments about reasons for valenced attitudes. I argue in chapter 2 that views that put ethical judgments before judgments about reasons seem to get things wrong about the role played by the latter in regulating our responses. On the other hand, identifying ethical judgments with judgments about reasons illuminates how ethical judgments regulate our responses by subsuming this phenomenon under the more general phenomenon of the regulation of responses by judgments about reasons.

I proceed to argue in chapter 3 that having reason to perform an act is a matter of the act's bringing about an and end that we should pursue, and that the idea of an end that we should pursue is simply that of an end that we should be motivated to pursue. Now as we have seen, valenced attitudes involve motivations – desiring a state of affairs involves motivation to bring it about, feeling obligated not to do something involves motivation not to do it, etc. Because ethical judgments are judgments that we should have valenced attitudes, and valenced attitudes involve motivations to act, ethical judgments entail that we should be motivated to act in certain ways. Hence, ethical judgments entail that we should be motivated in certain ways, and this in turn entails that we have reason to act in these ways. Good states of affairs are states of affairs we should desire and thus be moved to bring about, which means that they are states we have reason to bring about. Morally wrongful acts are acts we should feel obligated not to perform and thus be moved not to perform, which means that they are acts we have reason to bring about.

In chapters 3-5 I draw on this approach to the relationship between ethics, reasons for valenced attitudes, and reasons for action to explain how more exactly ethics bears on what we should do, and to solve various of the puzzles sketched in section 1.5. On my account, having reason to do something as an end in itself is a matter of our having reason to be intrinsically motivated to do it, or motivated to do it independent of its further consequences. I argue that the ethical status of many things is a matter of our having reason to have intrinsic pro- and con-attitudes towards them, or attitudes that involve motivations to do things independent of their further consequences. For instance, an act's intrinsic wrongness is a matter of our having reason to feel a kind of intrinsic guilt-tinged aversion towards performing it, and a state of affairs' intrinsic goodness is a

matter of our having reason to prefer it in itself rather than on account of what else it will lead to. As such, our reasons to avoid doing things that are intrinsically wrong and our reasons to do what brings about good states of affairs will be reasons to perform these acts as ends in themselves. I seek to show that all kinds of ethical categories are constituted by our reasons to have intrinsic pro- and con-attitudes. For instance, an act's intrinsic lowliness is a matter of our having reason to have a kind of intrinsic shame-tinged aversion towards performing it, an object or activity's intrinsic value is a matter of our having reason to have intrinsic appreciation for it, and an event's grievousness is a matter of our having reason to mourn it on its own account. By commanding that we should in these ways be moved to do things on their own account, these other ethical categories have a claim on our intrinsic reasons for action, or what we should do in and of itself.

An important aspect of my theory of the relationship between what we should feel and what we should do concerns the distinction between having reason to be somewhat motivated to do something and having reason to be motivated to do it on balance. Thus, you might be somewhat motivated to keep your leg, but motivated on balance to get rid of it if this is the only way to save your life. Decisive reasons to be somewhat motivated to do something are reasons to do it, but they can be outweighed by decisive reasons to be even more strongly motivated to do something else. Thus, it might be irrational not to want to keep your leg, but also irrational not to have a stronger desire to keep your life, in which case it would be irrational to keep your leg at the cost of your life. What we have most reason to do is what we should be motivated on balance to do. I argue, however, that certain motives are distinctive in that our reasons to have them at all can only be decisive if we should be moved by them on balance. Unlike desires to keep your leg, it seems that feelings of obligation to do something or motivations to do it out of a sense of honor cannot be rationally mandatory if it is rationally mandatory to be swayed by stronger motives to the contrary. If you have most reason to break a promise, it does not seem that you have to feel obligated to keep it. If you have most reason to plead for mercy, it does not seem that you have to feel moved by a sense of honor not to do so. I contend, however, that an act's moral wrongness consists in its being rationally mandatory to feel obligated not to do it, and that an act's lowliness consists in its being

rationally mandatory to be moved by a sense of honor not to do it.[45] Since these rational mandates require the absence of stronger reasons to perform the act, an act can only count as wrong or lowly if we have reason conclusive reason not to perform it.

Chapters 2-5 thus explain both ethics and facts about what we ultimately have reason to do in terms of facts about the pro- and con-attitudes we should take towards different things. In the final chapter of this dissertation I turn to the question of what it is to judge that we should have an attitude, which, given what has come before, is the question on which the complete analysis of ethics and reasons ultimately hinges. In this chapter I develop and defend a particular theory of what makes it the case that we should have a given response. I begin by reviewing a well known problem for analyses of ethical judgments in terms of reasons for attitudes of the kind I defend, which is to explain the kind of reasons we are talking about. Suppose, for instance, that a powerful evil demon who can read your mind threatens to harm your loved ones unless you desire that all cups in the universe are blue. This might seem to give you a kind of reason to desire that all cups are blue, but it would not make it good that all cups are blue. Now when we say that good states of affairs are states we have reason to desire, what we mean is that they are states towards which desire is warranted, appropriate, or fitting. On the other hand, the demon's threat does not make it the case that states in which all cups are blue warrant or befit desire – it merely makes it the case that we should do whatever we can to make it the case that we desire the state. But what distinguishes thinking that a consideration contributes to an attitude's fittingness from thinking that it merely counts in favor of getting oneself to have it?

One distinguishing feature is, for example, that thinking that a consideration contributes to the fittingness of a desire for a state of affairs is thinking that the consideration contributes to the goodness of the state of affairs. But those who seek to analyze good states of affairs in terms of fitting attitudes cannot use this fact as an *explanation* of the difference between thoughts about fittingness and thoughts about other reasons, on pain of vicious circularity. There have been several alternative attempts in

[45] More accurately: an act's wrongness or lowliness consist in its being rationally mandatory to have these attitudes unless one is going to refrain from performing the action anyway. This sort of conditional form of rational requirement is discussed in detail in Chapters 4 and 5.

the literature to explain the difference between considerations of fittingness and other considerations in terms of the content of such considerations, none of which seem satisfactory. I argue that a superior approach is to look to the mental states involved in judging that one has these different kinds of reasons. I point out that an intuitive difference is that judging an attitude fitting has a propensity to cause one to have it that is direct and unmediated by actions undertaken to get oneself to have it. Judging that you should desire a state of affairs on account of, say, its having a more equal distribution of income is capable of causing you to desire it directly, whereas judging you should desire a state on account of the fact that the demon will harm your loved ones if you don't cannot cause such a desire without your doing something (e.g. taking pills, classically conditioning yourself, etc.) to bring it about that you have it. In this way the distinction between fittingness and non-fittingness reasons to have valenced attitudes is exactly analogous to that between epistemic and non-epistemic reasons for belief. Judging that you shouldn't believe in God on account of, say, his explanatory superfluity can directly cause you to believe he doesn't exist, whereas merely judging that you shouldn't believe in God because it's much less depressing not to have to think that a manifestly cruel being is running the show cannot cause disbelief without first motivating you do things to bring it about.

I contend that the best way to capture this intuitive difference is to understand judgments about the fittingness of an attitude in terms of the acceptance of norms that prescribe having it. I survey two ways in which this can be done, corresponding to two different senses in which we can talk about someone's 'accepting a norm' that prescribes a response. To accept a norm that prescribes an attitude in what I call the 'shallow sense' is just to be in a state that has the functional roles of directly causing you to have the attitude, engendering a kind of cognitive dissonance if you don't, and combining with other states of its kind to form a "system of norms" that regulates your responses in more complex, structured ways.[46] When you accept a norm in this sense, your responses are

[46] This is, I think, the most plausible way of rendering the kind of state of norm acceptance discussed by Gibbard (1990). Gibbard's own approach to characterizing the state was different in some significant respects because it tied the notion of norm acceptance to the idea of normative discussion the adaptive function of which was coordination via consensus formation. But for many attitudes (e.g. beliefs and prudential motives) it is highly implausible to think that the primary adaptive function of their regulation by norms had anything to do with coordination via consensus formation.

directed by what the norm actually prescribes rather than what you might think it prescribes. On the other hand, to accept a norm that prescribes an attitude in what I call the 'deep sense' is to be in a state, representations of which have the role of directly causing you to have the attitudes that they represent the norm as prescribing. If we say things about people to the effect that they accept norms that prescribe certain responses, but fail to have those responses because they misunderstand the norms they accept, we are talking about norm acceptance in the deep sense.

One way to explain what it is for someone to judge an attitude fitting is just to identify the judgment with one's *shallow acceptance* of norms that prescribe it. An alternative is to identify one's judging an attitude is fitting with one's *representing* one's *deeply accepted* norms as prescribing the attitude. I argue that we should take the second approach on the grounds that it provides the best explanation, not only of how normative judgments guide our responses, but also what we are doing when we engage in basic normative inquiry. A significant and rather diverse group of philosophers have been attracted to the idea that basic normative inquiry uses reflective equilibrium methods to discover our own deepest "values" or "commitments."[47] On what I argue is the best version of this view, our basic normative inquiries into which attitudes are fitting is inquiry into what our deeply accepted norms prescribe. I contend that this makes best sense of the synthetic *a priori* nature of basic normative inquiry and its central features, including how normative intuitions of all levels of generality can provide direct evidence about the normative facts, how certain intuitions can lead us astray, and how seeking a unified theory of our non-debunked intuitions can enable us to hook onto the normative facts in much the way unified theories of perceptual experience enable us to hook onto facts about the external world.

I proceed to argue that the explanation of normative guidance and inquiry developed in terms of the deep acceptance of norms best supports a theory of normative judgment that I call *Norm Descriptivism*. According to this theory, judgments that an agent has reason to have a certain response are representations of the fact that the most fundamental norms she accepts prescribe that she have it. In the case of judgments about

[47] See for instance (Goodman 1954), (Rawls 1971), (M.B.E. Smith 1977, 1979), (Fischer and Ravizza 1992), (Kamm 1993), (Unger 1996), and (McMahan 2000).

one's own reasons, one represents the prescriptions of one's fundamental norms under a deliberative mode of presentation, or the mode of presentation under which one represents their prescriptions when one deliberates about how to respond. I argue that this mode of presentation is derived from the causal-explanatory relationships between one's normative intuitions and the states of norm acceptance they represent. In the case of judgments about other agents' reasons, one represents the prescriptions of their norms under a mode of presentation derived from their deliberations; roughly that of "whatever answers the questions they ask when they deliberate about how to respond." I contend, however, that shared normative inquiry standardly presupposes that the agents engaged in it accept the same fundamental norms, and that we typically make this assumption when we reason together. I adduce considerations from our experience with interpersonal normative discussion and evolutionary theory that seem to vindicate our presupposition that we all pretty much do accept the same fundamental norms. But I also draw upon these features of shared discussion and actual normative life to debunk the view that our normative judgments must presuppose that a reason for one agent is a reason for any conceivable agent placed in the same circumstances.

I argue that Norm Descriptivism provides a better explanation our normative thought than the views that I take to be its main competitors, both of which can be motivated by the arguments in favor of analyzing normative judgments in terms of the acceptance of norms. The first is *Norm Expressivism*, which (roughly) identifies judging that an agent should respond in a certain way with (shallowly) accepting a system of norms that prescribe responding in that way in her circumstances. The second is *Norm Relativism*, which identifies judging that an agent should respond in a certain way with representing the most fundamental norms one (deeply) accepts as prescribing responding in that way in her circumstances. I contend that Norm Expressivism cannot make adequate sense of what goes on in normative inquiry. I seek to show that Norm Relativism (and perhaps also Norm Expressivism) cannot account for our thinking about the epistemic access each agent has to her reasons, and that the Relativist's story of what we are doing when we talk to each other about how to respond cannot account for its symmetry with what we are doing when we deliberate about how to respond on our own. And I argue against both Norm Expressivism and Relativism on the grounds that they

cannot adequately account for the conceptual constraints on which beings we can hold subject to reasons, the relationship between how to respond and how one can respond, and the connection between being subject to reasons and being able to represent those reasons. I also defend Norm Descriptivism's implication that it is incoherent to think that an agent has reason to have a response but that her fundamental norms do not prescribe it in terms of a lack of appreciation of what distinguishes an agent's having reason to have a response from our having reason to want, hope, or make it the case that she has it.

I conclude by considering the implications of the Norm Descriptivist conclusions of the second section of the dissertation for the understanding of the relationship between ethics and fitting attitudes developed in dissertation's first section. If Norm Descriptivism is correct, then the fact that it is fitting for an agent to have an attitude is identical to the fact that the most fundamental norms she accepts prescribe that she have it. But most of the first section's analyses of ethical facts into facts about fitting attitudes spoke simply of the attitudes it is fitting "to have" without specifying for whom these attitudes are fitting. Of course, if we are correct in our presupposition that we all pretty much accept the same fundamental norms (as I argue we are), the same attitudes will be fitting on the part of pretty much all of us pretty much all of the time. But the fact that it is conceptually possible for agents' reasons for attitudes to differ demands that we account for whose reasons we are talking about when we analyze ethical concepts in terms of fitting attitudes.

My solution to this problem is to broaden the application of a contextualist strategy which has been suggested by Gibbard (1998) in the context of explaining whose reasons for preference we are talking about when we talk about good states of affairs *simpliciter*. When we say that an act is, say, blameworthy, we are talking about the reasons of a contextually salient set of agents to feel indignant. Even if there were space aliens who would not share these reasons, and even if there are sociopaths who do not actually share them, we could still talk truthfully about blameworthiness in a context that excluded such agents. These kinds of contextualist considerations also put constraints on our attributions of wrongness and lowliness via their connection to the concepts of blameworthiness and shamefulness. There could be space-aliens whose fundamental norms prescribe feeling obligated to torture us earthlings and prescribe feeling moved by

a sense of honor to slack off. But surely it would be false of us to say that the space aliens do wrong when they fail to torture us or that their hard work and discipline are lowly. But the Norm Descriptivist can account for this by explaining that to judge someone's act wrongful or lowly is not simply to judge it to be an act that the person should omit out of feelings of obligation or motivation by her sense of honor. It is moreover to judge it to be an act that it would be blameworthy or shameful to omit barring excuse. If we were to call an alien's act wrongful or lowly, we would imply that we should feel indignant or scornful towards her should she perform the act under conditions of full responsibility. But since we plainly would not have reason to be indignant with aliens for failing to torture us, or scornful of aliens who work hard, it would be false for us to call such acts wrongful or lowly - even if it would be true for us to say that they should feel obligated and moved by a sense of honor to perform these acts.

Chapter 2

From Judgmentalism to Fitting Attitude Analyses

Ethical judgments, recall, entail judgments about the kinds of motivationally laden pro- and con-attitudes we should have. That one has done something blameworthy entails that one should feel guilt, that a state of affairs is good entails that we should desire, wish, or be glad that it obtains, and so on. There might be a sense in which we could talk about having "reason to have attitudes" like guilt and desire if, say, a demon will harm our families if he detects that we don't have them. But the reasons for attitudes we take ourselves to have in making ethical judgments are of a different kind. In thinking that you have done something blameworthy, you think there are considerations which make it *fitting* or *appropriate* for you to feel guilt – like that you broke a promise because you were too lazy to keep it, or that you assaulted an innocent person for fun. Similarly, in thinking an outcome good you think that there are features of the outcome that make it fitting or appropriate to desire, wish, or be glad that it obtains – like that it would be an outcome in which more puppies are happy, or that it would be one in which income is distributed more equally.

The distinction between thinking a pro- or con-attitude fitting and thinking merely that you have reason to get yourself to have it is, as I mentioned, exactly analogous to the distinction between epistemic and pragmatic reasons for belief. One important thing that distinguishes thinking that a consideration contributes to a pro-attitude's fittingness or a belief's epistemic justification is the propensity of the thought to cause you to have the attitude or belief. Ordinarily, you do not need to *try* to get yourself to feel guilt if you realize that guilt would be fitting due to your breaking a promise for no good reason – the mere realization that you should feel guilt for having done this is sufficient to get you

feeling guilty. On the other hand, merely thinking that you "should feel" guilt (or that you should get yourself to feel guilt) because bad things will happen if you don't is not capable of directly causing you to have it. For this kind of thought to cause you to feel guilt, you (or some process within you) must do something in order to get yourself to feel it - like take a mind-altering substance, condition yourself, talk yourself into thinking guilt fitting, etc. In the same way, thinking that you should believe a theory because it is the simplest one can cause you to believe it without your having to do anything to instill the belief. But thinking that you should believe a theory because believing it will lift you out of your depression can do no such direct work. For this thought to cause you to believe the theory, it must first cause you (or some process within you) to do something in order to bring the belief about – like ignore certain bits of evidence and focus your attention on others.

There are two broad approaches we could take to explaining the relationship between ethical judgments and the judgments they entail about the fittingness of certain pro- and con-attitudes. First, we might think that ethical judgments are the more fundamental kinds of judgments in that we can understand them independently of judgments about which emotions and desires are fitting. If so, we could try to explain our judgments about the fittingness of emotions and desires in terms of our making ethical judgments that entail them. Thus, we could try explaining the judgment that it is fitting to feel guilt as a judgment that we have done something blameworthy, we could try explaining the judgment that it is fitting to desire a state of affairs as a judgment that the state of affairs is good, and so on. This order of explanation is typical of views which hold that pro- and con-attitudes themselves involve ethical representations, the most obvious being judgmentalism, according to which emotions and desires involve ethical judgments.

The alternative approach, which I favor, takes judgments about which kinds of emotions and desires are fitting as the more fundamental kinds of judgments. Proponents of this approach think that judgments about fitting attitudes can be understood prior to ethical judgments, and that we can explain our ethical judgments as judgments about the fittingness of emotions and desires. Thus, we could try explaining the judgment that we have done something wrong or blameworthy as a judgment that it is fitting for us to feel

guilt, we could try explaining the judgment that a state of affairs is good as a judgment that it is fitting for us to desire it, and so on. These kinds of explanations are analyses of ethical judgments in terms of fitting attitudes, which are often called *fitting attitude analyses*, or *FA-analyses* for short.

In this chapter I present a case for preferring the order of explanation offered by fitting attitude analyses. I point out several problems for views that take ethical judgments to be prior to fittingness assessments. These problems suggest that the psychological roles of fittingness assessments are unlike anything that must be involved in pro- and con-attitudes. They also suggest that the order of explanation provided by FA-analyses is needed to explain what ethical judgments are all about and how they govern our valenced attitudes. I seek to show how analyses of ethics in terms of fitting attitudes help to make sense of our ethical thought, and how they offer us an attractive way to subsume the central normative features of ethical judgments under the more general features of judgments about the justification or warrant of attitudes, which include judgments about epistemic reasons for belief.

2.1. Judgmentalism About Valenced Attitudes

The view I shall call *judgmentalism about valenced attitudes* is the view that part of what it is to have a pro- or con-attitude is to make an ethical judgment. The ethical judgment held to be involved in having the valenced attitude is the one that entails that the valenced attitude is fitting. This view thus encompasses what is referred to as 'judgmentalism about emotions' (or sometimes 'cognitivism about emotions'), which holds that feeling emotions like guilt involves making ethical judgments like that to the effect that one has done something blameworthy. But since the states usually referred to as emotions are a proper subset of those I am calling valenced attitudes, the thesis that I am discussing is broader. Judgmentalism about valenced attitudes includes views like that to the effect

that part of what it is to desire that a state of affairs obtains is to judge that the state of affairs is good, which has sometimes been referred to as 'the guise of the good thesis'.[48]

Although this kind of judgmentalism has to a certain extent fallen from grace, close descendants of the view are still something of an orthodoxy among philosophers, at least when it comes to understanding emotions as states that have a certain kind of ethical content. One source of intuitive attraction to this sort of view may be the way we tend to resort to ethical language in order to describe the phenomenology of pro- and con-attitudes. If asked what it feels like to feel angry at someone, to feel guilt for what we have done, to desire an outcome, or to feel ashamed, we may naturally reply that it feels "like the person has done something outrageous," "like we have done something wrong," "like the outcome would be good," and "like we have done something lowly." Indeed, some have argued that different constituent ethical judgments are needed to distinguish different emotions from each other.[49] The claim would be that we could not, for instance, distinguish or explain the difference between guilt and shame without saying that the former involved judgments about wrongness and the latter involved judgments about lowliness.

It is sometimes claimed that without judgmentalism (or something like it) we could not explain "the intentionality" of attitudes like emotions, but care must be taken about this claim. As Deigh (1994) observes, the view that emotions involve ethical judgments is in no way needed to explain the mere fact that emotions have intentional objects conceived of as people, acts, and states of affairs towards which they are directed. When you are angry *at* a person, feel guilt *for* what you did, or are frustrated *that* something is happening, your state of mind bears an intentional relation to the person, act, or state of affairs rather than the thoughts that the person has done wrong, that you have done wrong, or that the state of affairs is a frustrating one. Judgmentalists who portray their opponents as people who maintain that emotions are objectless states, or try (like Hume (1739) and James (1890)) to reduce such intentionality straightforwardly to eliciting conditions, are thus attacking straw men.

---

[48] For proponents of judgmentalism see for example Foot (1959) Kenny (1963), and Solomon (1976, 1988). For discussion and criticism of the guise of the good thesis, see for instance Velleman (1992).
[49] Such claims are advanced by Solomon (1976).

For considerations of the intentionality of emotions to count in favor of the judgmentalist position, judgmentalists must somehow try to show that their theory best accounts for *the kind of things that can be* the intentional objects of our emotions. Arguments that purported to show just this were provided by authors like Foot and Kenny, who claimed that it is conceptually impossible to have emotions towards things unless one has certain ethical beliefs about them. Thus, Foot maintained that:

> If a man does not hold the right beliefs [about something], then whatever his attitude is it is not pride. Consider, for instance, the suggestion that someone might be proud of the sky or the sea…This make sense only if a special assumption is made about his beliefs, for instance that he is under some crazy delusion and believes that he has saved the sky from falling, or the sea from drying up. The characteristic object of pride is something seen (a) as in some way a man's own, and (b) as some sort of achievement or advantage; without this object pride cannot be described (Foot 1959, 113-114).

Foot makes it clear that the sense in which she thinks one cannot feel pride without the right beliefs is that the pride is conceptually impossible, not simply unjustified: "I do not mean, of course, that one would be illogical in feeling pride towards something which one did not believe to be in some way splendid and in some way one's own, but that the concept of pride does not allow us to talk like that" (Foot 1963, 76). In a similar vein Kenny argued that quite generally it is conceptually impossible to have an emotion in the absence of a corresponding ethical belief:

> In fact, each of the emotions is appropriate – logically, and not just morally appropriate – only to certain objects…. What is not possible is to be grateful for, or proud of, something which one regards as evil unmixed with good….What is not possible…is to feel remorse for something in which one believes one had no part (Kenny 1963, 135).

One of the most important arguments in favor of judgmentalism is, however, its promise to explain how emotions and desires can be assessed as fitting and unfitting, and how these judgments about fittingness have so much in common with judgments about epistemic reasons for belief. For just about any state of mind, we can speak of its being the case that we "should [or shouldn't] have it" in the sense of its being a state that it's good or bad to be in, or a state that we should or shouldn't get ourselves into. In this kind of pragmatic or strategic sense we can say that we have reason not to have headaches, or

even tingling sensations – if, say, the evil demon will harm our loved ones if he detects that we do. But we cannot assess headaches and tingling sensations as states we should or shouldn't have in the sense of states that it is appropriate, warranted, or fitting to have in some situation. There is no kind of assessment of a headache or a tingle as a state we should be in that has a direct propensity to cause us to have it without our having to do anything to get ourselves to have it. Why, then, should states like emotions and desires be different from headaches and tingles in that we can assess them, not only as good or bad states to be in, but as states that can be fitting or unfitting? Why should emotions and desires, but not headaches and tingles, be capable of responding directly to our thoughts about a kind of reason to have them?

By way of an answer, the judgmentalist points us to one kind of mental state that definitely admits of both pragmatic and a non-pragmatic normative assessments – judgments, or beliefs. Like any other mental state, you can think that certain beliefs would be good or bad to have, and you can be moved by these thoughts to try to bring about the beliefs in yourself. But as we have seen, you can also think that you have epistemic reason to believe something, and this kind of thought can cause the belief without its having to motivate you to instill the belief in yourself. What the judgmentalist about valenced attitudes does is reduce fittingness assessments of pro- and con-attitudes to evidential assessments of the ethical judgments that these attitudes allegedly involve.[50] Like any other kind of judgment or belief, the judgment that you or someone else has done wrong, or that an outcome would be good, is directly sensitive to your views about whether it is supported by evidence. If valenced attitudes essentially involve such ethical judgments, they too will be directly sensitive to our views about the evidential status of these judgments. If taking away the judgment of blameworthiness *ipso facto* takes away the guilt, a power to directly destroy the judgment of blameworthiness is a power to directly destroy guilt. If taking away one's judgment that an outcome is good automatically takes away one's desire for that outcome, any assessment that can directly take away the former can directly take away the latter.

There are actually two kinds of judgments we can make about the fittingness of a valenced attitude, which come to largely the same thing when we are assessing our own,

[50] See for instance Foot (1959), Taylor (1985), and Greenspan (1988).

current attitudes. First, we can speak of an attitude's being *objectively fitting*, or the kind of attitude it would be fitting to have in a circumstance given all the (merely descriptive) facts about that circumstance - whether or not one has epistemic access to those facts. Thus, even if you have no evidence at all that a certain state of affairs will lead to the massive abatement of suffering at no cost whatsoever, the fact that it will do so can make it objectively fitting for you to desire that state of affairs. But second, we can speak of an attitude's being *subjectively fitting*, or the kind of attitude that it is fitting to have in a circumstance given the evidence you actually have in it. If, for instance, your evidence points overwhelmingly in favor of the view that I have just shot someone in cold blood, but unbeknownst to you we were just acting out a scene, it could be subjectively (but not objectively) fitting for you to be outraged at me.

According to the judgmentalist, to think that someone's emotion or desire is objectively fitting is to think that the ethical judgment it involves is true, while to think that someone's emotion or desire is subjectively fitting is to think that the person is epistemically justified in holding the ethical judgment she does. When you assess your own, current emotions and desires as fitting or unfitting, you will primarily be interested in whether they are objectively fitting. You want to know whether, say, you should desire a state, and you wonder whether it is actually good, or you want to know whether you are justified in being angry at someone and you wonder whether she actually did something culpably wrong. But your views about the objective fittingness of your attitudes are directly guided by your views about their subjective fittingness. You will tend to believe that a state is good or that someone has done culpable wrong just in case take your evidence to support the view that it is good or that she has done such wrong.[51] So your ethical judgments will respond directly to your views about whether you have epistemic reason to have them. If the judgmentalist is right that your emotions and

[51] Though as we have seen this tendency is not iron-clad. You might think that your evidence supports believing that certain things are good or culpable that you find yourself unable to believe are good or culpable. This might happen if, for instance, you got convinced that evidence in favor of Donald Marquis's view on the badness of death was manifestly overwhelming, and you were forced to concede that your evidence supported believing that embryo adoption was a very good thing and women who use the morning-after pill are among the most culpable killers. Of course, you might well come around to the view; my point is simply that you might (for awhile) both think you that you should believe the view but find yourself unable to believe something so counter-intuitive, much as some of us might be inclined to agree that we should believe the many-worlds interpretation of quantum mechanics but find ourselves unable to do so.

desires involve ethical judgments, these valenced attitudes too will, to a certain extent, be directly guided by assessments of your evidential reasons to hold ethical judgments.

2.2. Bloodless Judgments

Now as plausible as it might be to claim that pro- and con-attitudes involve ethical judgments, it does not seem very plausible to claim that they are *nothing more than* ethical judgments. We seem capable of what we might call *bloodless* ethical judgments, or ethical judgments that we make without actually having the corresponding pro-and con-attitudes. As we observed in Chapter 1, our ethical judgments appear to have a propensity to cause our valenced attitudes, but this propensity is not always decisive. It seems, for instance, that in a state of depression or mental exhaustion I might genuinely judge a state of affairs good, but have no actual desire for it to obtain. Similarly, consider someone who has been brought up in an environment in which doing a certain kind of thing (say living high and giving nothing to Oxfam) is encouraged, done by all, and thought so obviously permissible that it would be laughable to feel guilty about doing it. It seems that such a person could come to genuinely believe that it is deeply culpable to continue to do the thing in question without (especially at first) feeling any actual guilt for doing it.

What makes bloodless ethical judgments possible is the fact that pro- and con-attitudes essentially involve phenomenal, attention-directing, and motivational elements that ethical judgments can lack. Feeling guilt for having done something involves a certain felt quality that is unpleasant and akin to a kind of sinking sensation. It also involves motivation to do what can be described as "making amends" for what one has done –things that tend to restore the world to how it was before one did it, things that tend to be preferred by offended parties, and things that look like self-punishments. Guilt does not simply involve motivations to make amends when circumstances may arise. In addition to directing one's attention to what one has done and to what else one could have done, it tends to focus attention on the idea of making amends and motivates searching for opportunities to do so.

Similarly, it seems to be part of what it is to desire an outcome that one is motivated to bring it about if one can. But the kind of desire for a state of affairs the fittingness of which is entailed by its goodness is not simply a motivational state or a state that combines with representations of which acts will bring the state about to cause one to perform these acts.[52] To think a state of affairs good is not simply to think we should be moved to bring it about. For instance, it is presumably coherent to think that you should be moved to tell the truth or keep a promise even if you think that no good will come of it, and that the world would be a better place if you did otherwise. More precisely, to think one state of affairs better than another seems to involve more than simply thinking that we should be moved to bring about one rather than the other. Suppose, for instance, that we promised a dying mountaineer that we would make sure that his child gets educated.[53] It seems coherent to think that we should be moved to bring about the education of his child at the expense of two others, even as we think that the world would be a better place if we broke our promise and brought about the education of the two children at the expense of the mountaineer's child.

What is it, then, to desire or prefer a state in the relevant sense? For one thing, it seems to involve directing one's attention to features of the state, and to whether there are ways in which the state can be brought about. Thus, as Darwall and Scanlon observe:

> A desire consists not simply in the capacity to be moved by awareness of facts regarding its object. It is both more active and more focused than that. It includes dispositions to *think* about its object, to *inquire* into whether there are conditions that enable its realization (Darwall 1983, 40-41).

> A person has a desire in the directed-attention sense that *P* if the thought of *P* keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favor of *P*…This idea seems to me to capture an essential element in the intuitive notion of (occurrent) desire….this character is generally missing in some cases in which we say that a person who does something for a reason nonetheless "has no desire to do it," as when, for example, one must tell a friend some unwelcome news (Scanlon 1998, 39).

It is natural to follow Scanlon in speaking of the relevant kind of desire for a state of affairs as involving "seeing the state in a favorable light." While it is more speculative to

[52] For this characterization of motivations or 'desires' in a thin sense see for instance Stalnaker (1984).
[53] For this kind of case see Carritt (1947).

attribute a phenomenal element to desires than to emotions like guilt, I think it plausible that there is a distinctive way it feels to desire or prefer a state in the relevant directed-attention sense.[54] There seems to be something it is like to see an outcomes one desires in a favorable light, which is similar to what it's like to "look forward to" the outcome. The thought of a desired state seems pleasant, and the thought of a preferred state occurring instead of a dispreferred one feels like "being relieved." Having a desire for an outcome may also involve feeling something that might be metaphorically described as a sort of "tug in the direction of the outcome."

Our depressive can thus bloodlessly think a state of affairs good if she thinks it good but lacks either the motivation to bring it about, or the tendencies to focus on it and how she might bring it about, or the right kind of subjective experience involved in desiring it. And our recent convert can bloodlessly judge himself culpable for living high and letting die if he judges himself culpable but lacks either the unpleasant, sinking sensation, or the motivation to make amends, or the focus of attention on what he has done and how he can make up for it involved in feeling guilt for what has done. That there can be (and are) cases like this probably sounds like common sense to most of us. But it may be worth our while to consider the following challenge. Why think in such cases that the depressive actually *judges* that the state is good or that the recent convert actually *judges* himself culpable? If the depressive says the state is good but does not desire it or the convert says he is culpable but feels no guilt, why not say, as Prinz (2006) puts it, that these individuals are simply 'paying lip service' to the views they espouse without actually holding them?

The reason, it seems, is that ethical judgments like those about culpability and goodness seem to play certain psychological roles that are distinct form those of pro- and con-attitudes, and that they can continue to play them in the absence of these attitudes. First, like beliefs or judgments more generally, ethical judgments play a role in domain-general reasoning and inference. They can combine with other judgments, both ethical and non-ethical, as premises and conclusions in trains of thinking. If someone thinks that

[54] While some philosophers of mind claim that desire has a distinctive phenomenal element, others tend to find this view implausible. Part of the reason for this may be that the first group of people have in mind 'desire' in the directed-attention sense, while the second group has in mind 'desire' in the sense of any (possibly domain general) motivational state whatsoever. Another complication may concern as the distinction between occurrent and non-occurrent attitudes, a topic to which we shall come in time).

his living high and letting die is culpably wrong, he will have tendencies to draw inferences that he would lack if he did not think his behavior culpably wrong. He would, for instance, have a tendency to infer that similar behavior by others is similarly culpably wrong, and these judgments might themselves lead to attitudes like anger at others for failing to give to the poor. He might also have a tendency to infer that others who seek to punish him or hold him accountable for his living high and letting die are within their rights to do so, and feel no anger at them for doing this. Those who do not think themselves culpable would not have the same inferential tendencies and corresponding tendencies to be angry at others for failing to give and to feel no anger towards those who behave punitively towards non-givers. But these tendencies to inference and subsequent pro- or con-attitudes (or absence thereof) can be present in the absence of any feeling of guilt for living high and letting die. In the same way, a person might conclude from Nozick's (1974) experience machine case that her own knowledge would be intrinsically good, but find herself with no intrinsic desire to know things. Yet she might infer that it is also intrinsically good for others to know things, and desire that they have such knowledge quite independently of its consequences. It might be one thing to have an even more glowing attitude towards the education of the underprivileged on account of thinking that it has non-instrumental value, but quite another to have a glowing attitude to knowing things that one finds boring oneself.

In a related way, ethical judgments seem to play a role in shaping our views about what we and others should do, which can alter our behavior without our actually coming to have the emotions and desires that the ethical judgments license. The above tendencies to infer that others like one are blameworthy or those who punish us are within their rights can of course make a difference to our behavior – one may behave punitively towards others who do not give, and may offer no resistance to whose who behave punitively towards one on account of one's failures to do so. Moreover, the judgment that one has done something blameworthy seems capable of causing one to think that one should make amends for what one has done and causing one to make amends as a result – all without one's actually feeling any guilt. One would, in such cases, be motivated to behave in ways that one who feels guilt would behave, but one might still lack guilt on account of lacking the right phenomenology and tendencies to

direct attention. In the same way, the judgment that knowledge is good might cause one to think that one should pursue knowledge and cause one to actually pursue knowledge (rather grudgingly) as a result – all without one's actually desiring (in the directed-attention sense) to know.

Finally, ethical judgments seem to play a role in the ways we feel and think about ourselves and others in light of having or failing to have the pro- and con-attitudes that the ethical judgments license or mandate. Judging that you have done culpable wrong but feeling no guilt is a state of failing to feel what one thinks one should. Such states tend to engender thoughts to the effect that you are somehow defective, or that there is "something wrong with you" for failing to feel as you ought, which can come out in other attitudes or behavior. You might, for instance, tend to direct attention in ways geared towards instilling the attitude you think you should have. In this way the recent convert to the view that he is culpable for living high and letting die might review the evidence in favor of his position, attempt to vividly appreciate the features in virtue of which he thinks himself culpable, relate this information to more emotionally charged personal experiences (like episodes in which he or his loved ones were denied help that they needed), and so on. In certain respects this might resemble the patterns of directed attention that go with feeling guilt for what one has done, but (i) the judgment of culpability might cause such patterns of attention in the absence of the phenomenology (or sometimes the motivations) associated with feeling guilt, and (ii) unlike the patterns of attention involved in feeling guilt, these patterns look more concerned with the features one takes to be culpability-making and one's appreciation of them and less concerned with what else one could have done and what one can do to make amends.

Failing to feel what you think you should also tends to engender con-attitudes towards oneself (shame, for instance), and may lead one to upbraid oneself in public or private for one's perceived failing. This is distinct from actually feeling guilt for what one has done in that the object of one's negative feeling and self-punishment is one's failure (or oneself on account of one's failure) to feel guilt for what one has done rather than what one has done itself. In more extreme cases, failing to feel what you think you should might motivate attempts to more directly instill the feelings in yourself. You might seek to instill the attitudes that you take to be fitting by the same means by which

you might try to instill an attitude for pragmatic reasons – therapy, conditioning, and drugs, for instance. But you might strive to do this simply on account of the attitudes' fittingness and quite independently of their benefits. You might try to reinforce your tendency to feel what you should simply as a way of mending the defect or thing wrong with you that is constituted by your failing to feel what you should.

2.3. Recalcitrant Attitudes

It seems, then, that judgmentalists cannot simply identify pro- and con-attitudes with the ethical judgments that entail their fittingness. Most plausibly, judgmentalists may claim that part (but not all) of what it is to have attitudes like emotions and desires is to make the corresponding ethical judgments.[55] But even judgmentalism so understood appears to face a problem. We seem to be capable of having emotions and desires that we judge to be unfitting or inappropriate. You can, for instance, feel guilt in spite of the fact that you believe that you have done nothing blameworthy. If you called someone to come to you, and in walking to you he steps on a landmine that has been secretly planted by a madman, you might well feel guilt for causing that person's death, even though you are confident that, as you had no evidence that anything was amiss, you did nothing blameworthy.

Similarly, you can desire an outcome that you actually think is bad, or prefer one outcome to an alternative that you think is better. Upon encountering someone trapped in a machine that will crush him, you might find yourself with a strong desire to allow him to be crushed out of a kind of morbid curiosity to "see how flat a person can be."[56]

---

[55] An alternative might be to claim that pro- and con-attitudes are special kinds of ethical judgments – the kinds that represent (or signify) ethical content under a distinctively phenomenal, motivational, and attention-directing mode of presentation. This alternative seems somewhat more complicated and is, I believe, rather less plausible. If we know that ethical judgments can occur without the relevant phenomenology, motivation, and attention-direction, and that they tend to play much the same inferential, motivational, and evaluative roles whether or not they are so accompanied, why should we say we have one kind of judgment when these other elements are present and quite another kind of judgment when they are absent? Since most of the same points that I shall make about the "judgment-as-proper-component" version of judgmentalism (and its quasi-judgmentalist descendants) carry over to the "special-mode-of-presentation" version, I shall for the sake of simplicity omit explicit discussion of the latter.

[56] This kind of case is drawn from an example of Malam (1989), building on an example of Tooley (1980). Another commonly given example of a desire of the relevant kind might be a spontaneous inclination to drink antifreeze or a can of paint.

Although you think that of course his being crushed would be horrible, and that of course it would be far better for him to be freed from the machine, you might find yourself with a kind of preference to crush him. You might feel attracted to the idea of seeing him crushed, find your attention directed to ways of bringing the crushing about, and find yourself somewhat moved by your preference to do what will bring about the crushing.

In a sense, then, Kenny (1963) seems to be flat-out wrong in his claim that it is "not possible to be grateful for, or proud of, something which one regards as evil unmixed with good." Suppose a man had been raised in a society where it was considered great entertainment to take sadistic pleasure in watching beatings of members of a despised underclass. It seems that the man could come to break with his culture, and strongly believe that watching such beatings is indeed evil unmixed with good – coming to think, for instance, not only that avoiding harms to the victims is more important than being entertained, but that sadistic pleasure derived from the actual suffering of others is in no way worth having. However, it would not be surprising, especially at first, if the man actually felt some gratitude towards people in his society who offered to take him to beatings or beat people in front of him. The man would presumably take his feelings of gratitude to be misplaced (especially if he could expect the other members of his society to know the errors of their ways), but he could have them nonetheless. In the same way, the culture in which the man was raised might have prized the sort of "toughness of mind" that enables one to kill and torture members of the underclass without "getting emotional" and without any remorse. Having broken with his culture, the man might now regard his ability to coolly kill and torture members of the underclass as a vice that is evil unmixed with good – the ability to maintain discipline under pressure is good, he thinks, but killing and torturing defenseless people is no test of courage. Yet, especially at first, it seems that the man might feel the stirrings of pride at thoughts of just how cold-blooded he can be to the underclass. He would take these feelings of pride to be completely inappropriate and unwarranted, but he could experience them nonetheless.

The above kinds of emotions and desires, which we have but take to be unfitting or inappropriate, are usually referred to as *recalcitrant attitudes*, as they remain in spite of the propensity of our normative judgments to remove them.[57] As with bloodless

[57] See for instance D'Arms and Jacobson (2003).

judgments, we might ask why we should take the apparent phenomenon of recalcitrant attitudes at face value. When someone claims that she has done nothing blameworthy but feels guilt, or claims that a state of affairs is bad but desires it, why not say that she is simply "paying lip service" to these views without actually holding them? The answer, it seems, is very much like the answer to the analogous question as to why we should think that bloodless judgments are possible. Judgments that attitudes are unfitting can play the same important inferential, action-guiding, and self-assessing roles in the presence of the attitudes that judgments about an attitude's fittingness can play in its absence.

You might experience the phenomenology, motivations, and directed attention associated with feeling guilt for calling the person to you who was killed by the landmine. But your judgment that you had done nothing blameworthy would involve tendencies to infer that others who do the same as you are not to blame, and explain your lack of anger at them. In the same way, you might judge that others are *not* within their rights to be angry at you or punish you for the death of the person, and cause you to resent their anger or attempts to hold you accountable for the person's death. Your judgment that you have done nothing culpably wrong might check your tendencies to try to make amends to the victim's family or engage in self-punishing acts. And your judgment that your guilt is inappropriate might lead you to try to expiate it – to focus on the features that make it unfitting and on how you wouldn't blame anyone for what you did, and (if the guilt is serious enough) to seek therapy or other means of purging the guilt. So too, our man from the culture with the underclass might have the subjective experience, motivation, and patterns of attention involved in desiring to beat members of that underclass. But his judgment that it would be a bad thing for him to beat these people might cause him to infer that it is bad thing for anyone else to do it, desire that the practice be stamped out, and contribute to organizations that promise to end it. His judgment that it would be a bad thing to beat members of the underclass might similarly check his desire to beat them himself, and might motivate him to try to expiate this desire by means of attention direction and therapy if need be.

Now, judgmentalism doesn't directly rule out the possibility of recalcitrant pro- and con-attitudes. What judgmentalism entails is that if you do have a recalcitrant

valenced attitude, you must be making conflicting ethical judgments – one that entails that the attitude is unfitting, and one that is involved in having the attitude itself. If feeling guilt for doing something involves judging oneself blameworthy for doing it, then feeling guilt for doing something that one doesn't think blameworthy involves both judging that one's act was blameworthy and judging that it wasn't. If preferring a state of affairs involves judging that it is better than its alternatives, then preferring a state to alternatives that one judges to be better involves judging the state better and judging it worse.

The problem for judgmentalism is not that we cannot make sense of what it is to hold conflicting beliefs or to make conflicting judgments. The problem is that merely having recalcitrant attitudes does not seem to have to involve the essential features of a conflict in ethical judgment.[58] As we have seen, ethical judgments involve characteristic tendencies to draw certain inferences, make certain judgments about what to do, perform certain actions, and have certain attitudes towards oneself on account of having or failing to have the corresponding pro- and con-attitudes. Conflicting ethical judgments would thus involve conflicting tendencies of these kinds. It seems, however, that we can have recalcitrant emotions and desires without any such conflicts in the inferences we tend to draw, the decisions we tend to make, or the evaluations we tend to have of our attitudes.

Suppose, for example, that I feel guilt for knocking over and breaking a friend's lamp, though I exercised all due caution and think that I did nothing that was culpably wrong or blameworthy. Perhaps I was pushed into the lamp, or someone put the lamp in my path when I wasn't looking. Were I really conflicted in my judgments about whether my behavior was blameworthy, it seems that I would have to have something like the following:

[58] Deigh (1994, 837) makes this kind of observation, as do D'Arms and Jacobson (2003, 129-130), who discuss the difference between merely fearing flying and moreover judging flying dangerous, and the relevance of this to the judgmentalist's need to posit inconsistent judgments wherever there are recalcitrant emotions. They discuss how those with phobic fears of flying "are typically well aware that [flying] is safer than activities they do not fear, such as driving to the airport…they do not worry when their friends fly, or buy insurance when forced to fly themselves," concluding in their footnote 7 that "The great challenge for judgmentalist accounts of recalcitrant emotion is that the behavioral evidence supporting the attribution of the evidentially suspect belief is problematic."

(J1) Conflicting tendencies to draw inferences about the moral status of similar and related behavior. These would include things like conflicting tendencies to judge that other lamp breakers in the same circumstances as I are blameworthy (and perhaps conflicting tendencies to get angry with them and to not get angry at them), and conflicting tendencies to judge that others are within their rights to be angry with me and hold me to account by scolding me (which might be manifest in conflicting tendencies to - on the one hand - get angry at such anger and scolding and - on the other - to accept it).

(J2) Conflicting tendencies to draw inferences about what to do in light of what I have done, and to be motivated to act accordingly. These would include tendencies to hold conflicting views about whether I should continue to apologize for the broken lamp and whether I should try to pay for repairs or a replacement, and correspondingly tendencies to have conflicting motives to do or abstain from doing what the guilt I feel motivates me to do.

(J3) Conflicting tendencies to evaluate my guilt in a positive or negative light. These would include conflicting tendencies to view my guilt as appropriate or inappropriate, and corresponding tendencies to think that something is going wrong with me in feeling the guilt I feel. They could also include conflicting tendencies to either retain the guilt or expiate it by means of attention direction and, should the need and opportunity arise, by such means as therapy and mind-altering substances.

It seems, however, that I could feel guilt about breaking the lamp and judge that I have done nothing blameworthy without any of these kinds of conflicts. It appears that I could feel guilt for breaking the lamp, with its characteristic phenomenology, motivations, and focused attention, yet manifest only those tendencies to inference, decision, and emotional evaluation associated with judging myself blameless. I could infer unambiguously that other lamp breakers like myself are innocent, and that others are unjustified in resenting my behavior or scolding me. Accordingly, I could have no tendency to get angry at other lamp breakers, every tendency to get angry at those who resent or scold lamp breakers like myself, and no tendency to accept such anger or

scolding. I could infer unambiguously that I have no reason to apologize or offer to make restitution, and strive only to check (and not at all to indulge) the motives I have to do so. I could unambiguously view my guilt as inappropriate, think only that something has gone wrong with me for feeling it, and focus attention and therapeutic efforts only on the expiation of my guilt. I could, it seems, do all this and yet still feel guilt for breaking the lamp. Surely, then, we must conclude that I feel guilt for having done something that I simply do not judge to have been blameworthy.

In the same way, consider the man from the culture with the despised underclass who judges it "evil unmixed with good" to see members of the underclass beaten, but still finds himself with the desire or preference to observe such beatings. Were the man really conflicted in his judgments about whether seeing such beatings is good, it seems that he would have to manifest something like the following:

(J1) Conflicting tendencies to draw inferences about the value of related events. These would include things like conflicting tendencies to judge it good and to judge it bad that other members of his culture to get to witness the beatings (as manifest, perhaps, in desiring both that they get to see them and that they do not), and conflicting tendencies to judge it good and to judge it bad that others make it difficult for him to see the beatings (as might be manifest in desiring both that they do this and that they not do this).

(J2) Conflicting tendencies to draw inferences about whether to bring about his observation of the beatings and related events. These would include tendencies to hold conflicting views about whether to bring it about that others see the beatings and whether to help others set up obstacles to his seeing them. Tendencies to make such conflicting judgments about what to do would result in conflicting motivations to resist or give in to one's desire to see the beatings and conflicting motivations to help and to hinder organizations that try to prevent himself and others from observing the beatings.

(J3) Conflicting tendencies to evaluate his desire to see the beatings in a positive or negative light. These would include conflicting tendencies to view his desire as warranted or unwarranted, and corresponding tendencies to think that something is going

wrong with him for having the desire he does. These could also include conflicting motivations to either retain the desire or expiate it by means of attention direction, therapy, and drugs.

But it seems that the man could have a desire to see the beatings and judge them bad in every respect without any such conflicts. He could desire to see the beatings, and have the motivations, focused attention, and subjective experiences this involves, yet have only those tendencies to inference, decision, and desire-evaluation associated with judging it bad to see the beatings. He could tend to infer that it is a bad thing for others to watch the beatings and a good thing for someone to impede his seeing them, and have no tendency at all to judge it good for others to watch or bad for someone to impede his watching them. Consequently, he could tend to desire that others not watch the beatings and that someone stop his doing so, without having any tendency to desire that others watch them or that no one stop him from watching. He could tend to judge that he should prevent everyone from seeing the beatings and lack any tendency to judge that he should enable himself or others to see them. He might strive only to check his motives to see the beatings, and contribute unambiguously to the efforts of organizations trying to prevent their viewing. He could tend to view his desire to see the beatings as unwarranted, think something is wrong with him for having it, and try to expiate the desire, without the least tendency to have the views and action tendencies associated with thinking it warranted. The man could do all of this, and still have his desire to watch the beatings of the members of the underclass. In such a situation, we must concede that he desires to bring something about that he does not judge to be good.

The judgmentalist might try to object that the foregoing criteria for attributing conflicting ethical judgments are too stringent. When we attribute conflicting judgments or beliefs to someone, we do not always require that she have conflicting inferential and action tendencies all across the cognitive board. We can, it seems, allow for somewhat "inferentially encapsulated" beliefs, like those we might attribute to people who espouse traditional Christian beliefs in God, Heaven, Hell, and the rest of it, but almost always reason and behave like run-of-the-mill atheists. Why, the judgmentalist might ask, can't

we regard recalcitrant emotions as states that involve inferentially encapsulated ethical judgments, which have few tendencies in competition with the judgments that make them recalcitrant, but exist in conflict with them nonetheless?[59]

The answer is that while there might be judgments that are relatively encapsulated, we lack reason to believe in judgments that are utterly impotent. There must be some distinction between the person whose Christian beliefs are inferentially encapsulated, and the person who has no Christian beliefs at all but merely "pays lip service" to Christianity – who might think he holds Christian beliefs but would be mistaken to think so. The distinction seems to be that the person with inferentially encapsulated beliefs has *some* tendency to infer in accordance with and to be guided in action by his Christian beliefs. There must be some range of inferences that he is prepared to draw from the beliefs, or some range of actions or contexts of action in which he will tend to be guided by the beliefs - narrow though these ranges may be - for him to count as holding the beliefs at all. Perhaps it is only on Sunday, or only inferences that certain things will get you sent to hell, or only certain acts of prayer (not undertaken purely for mediation, participation in a tradition, play-acting, etc.). But without at least some such tendencies to inference and action, it seems that one could not count as genuinely holding Christian beliefs that conflict with the atheistic beliefs that guide the rest of one's inference and action.

The same criteria apply to determining when an ethical judgment is relatively encapsulated and when it does not exist at all. It seems, however, that one can have recalcitrant attitudes without any of the tendencies associated with the corresponding ethical judgments. The above examples seemed to show that one could feel guilt or desire a state – *not only* without *many* of the inferential and evaluative tendencies involved in thinking oneself blameworthy or thinking the state good – but *indeed* without *any* of these tendencies. It seemed that I could have sufficient phenomenology, motivations, and focused attention to count as feeling guilty for breaking the lamp without my tending to make even the narrowest range of inferences or evaluations associated with judging my behavior blameworthy. Similar remarks hold for the man

[59] I am grateful to Any Egan for drawing my attention to this issue.

who can desire to see beatings of the underclass without any range of inferential or evaluative tendencies associated with judging it good.

In my descriptions of how someone could feel recalcitrant guilt or have recalcitrant desires, I stipulated that a person lacks certain inferential and evaluative tendencies and observed that it seems quite possible for him to still have the relevant recalcitrant attitude. I surveyed those inferential and evaluative tendencies that I could think of as relevant to the ethical judgments in question. If a judgmentalist thinks that I have left something out, it seems to me that the burden of proof is on her to identify a constituent feature of judging oneself blameworthy or desiring a state of affairs, and to argue that without this feature one could not have the recalcitrant guilt or desire. One thing that it seems the judgmentalist cannot point to, however, is the bare fact that one feels the guilt or has the desire one does. The judgmentalist cannot simply count feeling guilt or desiring a state of affairs as sufficient for "a kind of ethical judgment," on pain of collapsing her view into triviality. We could, of course, call emotions and desires 'ethical judgments' in a sense that signified nothing of theoretical importance. But then we could not complain of circularity if people were to undertake to explain states that we might call 'ethical judgments' in a different sense - like those that in section 2.2. we saw we can have in the absence of the corresponding emotions and desires - in terms of emotions and desires themselves.[60]

[60] Like remarks go for something else that the judgmentalist might try to say in the opposite direction. She might simply refuse to share (or acknowledge her sharing) the above intuitions about cases of recalcitrant judgment by pounding the table and insisting that she shall not call syndromes of phenomenology, motivation, and attention-direction associated with attitudes like guilt and desire 'guilt' and 'desire' unless they involve the right ethical judgments. But she then cannot complain about circularity if someone tries to analyze ethical judgments in terms of the syndromes of phenomenology, motivation, and attention-direction that I have been discussing; she will only have grounds for complaint of circularity if that person insists on calling these states 'guilt' and 'desire' in the same sense as her. If the reader is a judgmentalist like this, I ask her to mentally replace my talk about 'guilt' and 'desire' with 'schmilt' and 'schmesire', and read everything else as is. I'd also ask her why she cares so much about using the terms 'guilt' and 'desire' as labels for non-recalictrant schmilt and schmesire. Wouldn't the whole thing be easier if you just don't quibble so much about what we use terms to express and save me the trouble of writing 'sch' a bunch of times?

2.4. Quasi-Judgmentalism

In light of the problems posed by the phenomenon of recalcitrant valenced attitudes, those with judgmentalist leanings sometimes opt what we might call *quasi-judgmentalism*. According to quasi-judgmentalism, pro- and con-attitudes involve some sort of "ethical evaluation," or state with ethical content, but this state with ethical content need not be an ethical judgment. On one version of this view, pro- and con-attitudes involve somehow *entertaining* the corresponding ethical thought, or "seeing things in terms of" the ethical concepts the thought involves, without necessarily affirming the thought.[61] Thus, the idea might go, in order to feel guilt for what one has done, one must entertain the thought that one's action was blameworthy, or somehow "construe" or "see" it as blameworthy. To desire a state of affairs, the thought must occur to you that it is good, or you must somehow construe or see it as good.

It is important that we have some idea as to what the quasi-judgmentalist means here by 'entertaining' an ethical thought or 'construing' or 'seeing' something in terms of an ethical category. Our previous discussion seems to show that one need not lend any serious credence to an ethical judgment in order to have the corresponding valenced attitude. You can, it seems, feel guilt for breaking a lamp or desire to see someone beaten while being fully confident in your judgment that you are blameless or that it would be evil unmixed with good to see it. Nor does it look like you need to be considering the thought that you are to blame or that the sight is good in something like the way you might consider a view for the sake of argument. Similarly, talk of your "construing / seeing your act as blameworthy" and "construing / seeing the state as good" had better be more than just a fancy way of saying that you feel guilt or that you desire the state if the quasi-judgmentalist view is going to be substantive and pose an objection to analyzing ethical concepts in terms of emotions and desires.

The quasi-judgmentalist idea seems to be that pro- and con-attitudes involve states that are more akin to "ethical perceptions" than ethical judgments. Although perceptual states may be distinct from thoughts, having certain perceptual states might

[61] For examples of quasi-judgmentalists treatments of this kind see for instance (Roberts 1988) and (Greenspan 1988). For criticisms of quasi-judgmentalism related to (as well as distinct from) those I present here see (Gibbard 1990, 39-40 and 129-132), and (D'Arms and Jacobson 2003).

depend upon deploying certain concepts. For instance, you can see the Duck-rabbit image as either a duck or a rabbit, and you can see the Rubin's vase image as either two faces or a vase:

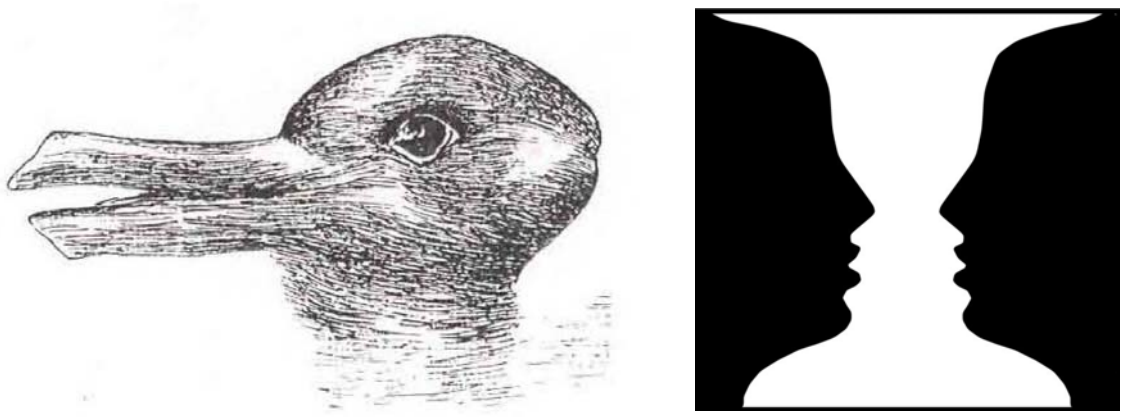

Figure 1: Duck-Rabbit and Rubin's Vase[62]

Psychologists explain our having different perceptual states when viewing such ambiguous images in terms of "top down" processing, which structures our sensory information in terms of different concepts from long-term memory.[63] Perceiving the Duck-rabbit as a duck or a rabbit depending upon whether one is deploying the concept DUCK or RABBIT to structure one's sensations is precisely the example used by Roberts (1988, 187-188) to explain what he means by "construing something in terms of a concept." The idea, then, might be that feeling guilt involves perceiving your action as blameworthy, and this involves somehow deploying the concept of blameworthiness to structure the information you have about your action.

Perceptual states are distinct from states of belief or judgment in several important respects. First, they are more "domain specific" in that they are less sensitive to learning and multiple sources of information than are beliefs. Thus, you can continue to perceive a stick in water as bent even if you have examined its shape and know the optics behind the illusion, and you can continue to perceive line (a) as longer than line (b) in the below depiction of the Müller-Lyer illusion even if you know the psychological story behind it and have used a ruler to see that the lines are of equal length.

[62] Rubin's Vase image taken from Braun (2000).
[63] Zimbardo and Weber (1997), Ch 5.

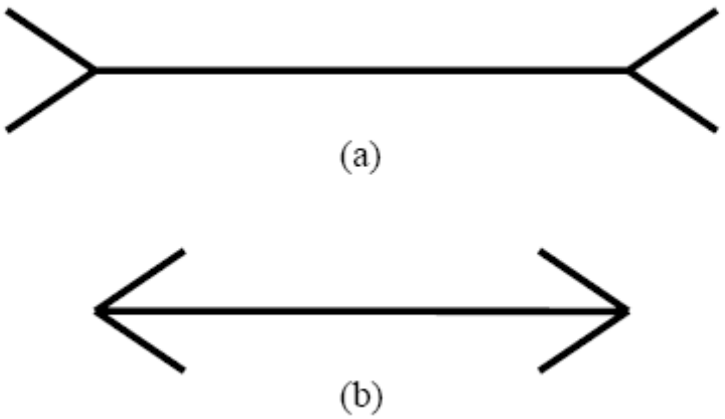

Figure 2. The Müller-Lyer illusion

What you perceive is sensitive to the visual information presented, and perhaps some "top-down" processing as ambiguous image perception suggests, but your perceptions are less sensitive than your beliefs to inferences from other things that you believe.

Perceptual states are also distinct from beliefs or judgments in that they can be more quickly instanced and can play roles in guiding behavior before beliefs have had a chance to come on the scene. Thus, you might be startled by your perception of something that looks like a snake or something that is approaching quickly before you have a chance to form the belief that a snake is present or that something is rapidly approaching you.[64] Another important feature of perceptual states related to their relatively rapid deployment is that they cannot be consciously inferred. Of course, such states are the result of sub-conscious processes that integrate a great deal of information, and in some cases we may perceive things as a causal consequence of certain consciously held beliefs (like that the artist who drew the image before us was trying to draw a duck). But the former informational integration process is not conscious, and the latter causal process is not inferential.

A final important feature of perceptual states, which is related to the fact that they are not inferred, is that they contribute their contents as "starting points" or data to processes of inquiry and belief formation. These starting points are by no means held to be infallible – they can of course be debunked by theories that best explain the totality of

[64] See for instance Damasio (1994) and LeDoux (1998).

such contents. It is simply that inquiry sets a burden of proof in favor of the contents of perceptual states; they are treated as veridical until proven illusory, and data to be fit by theory until theory has shown us how to explain them away. Debunking explanations of perceptual contents earn their own keep by offering a best explanation of all such contents together. In determining one's perception of the lengths of the Müller-Lyer lines to be mistaken by measuring them with a ruler, one relies on similar kinds of perceptions of how many notches each line takes up on the ruler, and in explaining the perception away one relies upon similar kinds of perceptions in assembling one's optical or psychological theory.[65]

To the extent that we have "sub-judgmental" ethical evaluations or appearances that are analogous to perceptual states, I think that we know them as "ethical intuitions." These states have the same relation to moral judgments that perceptual states have to empirical judgments and play analogous roles to those of perceptual representations. Such intuitions are rather more immune to revision in the face of contrary evidence than ethical judgments, more quickly deployed, un-inferred, and play the role of starting points or data in favor of which a burden of proof is set in ethical inquiry. Thus, someone raised in a racist culture might agree that all kinds of considerations decisively show that he owes the same to members of other races as he owes to members of his own. He might, as a result, genuinely judge himself to owe the same to people regardless of race, as manifest in his having the kinds of tendencies to inference, decision, and evaluation discussed in sections 2.2 and 2.3. But he might still have intuitive appearances to the effect that he owes more to members of his own race than he does to members of other races. In contexts in which he has no time to think, he might act on his racist intuitions, showing preference for members of his own race. He may continue to have un-inferred seemings that he owes more to "his own kind," even though he has dismissed them as unable to fit with the totality of the seemings he has about what he owes others. But even in constructing a debunking explanation of his racist seemings as, say, the product of mistaking the ethical relevance of certain special relations for the ethical relevance of race membership, the ex-racist would have had to rely on similar seemings about what he

[65] If, for instance, you thought you had stronger reason to think that the line-to-ruler appearances were more suspect than the line-to-line appearances, you might well construct a theory according to which the lines were not of equal length, and come to judge accordingly.

owes – like that skin-shade *per se* doesn't affect one's obligations, that one owes the same to adopted children as one does to biological children, and so on.

With these roles of sub-judgmental ethical evaluations or intuitions in mind, we should not be surprised that we can have pro- and con-attitudes that defy not only our ethical judgments but indeed our ethical intuitions. Not only can we have recalcitrant valenced attitudes that we judge to be unfitting, we can also have intuitively recalcitrant attitudes that we simply intuit to be unfitting (whether or not we judge that they are so). Suppose again that I feel guilt for knocking over my friend's lamp, but instead of judging myself to be blameless I suspended judgment – I am unsure, let's say, about what to infer about similar cases, about the appropriateness of others' anger with me and scolding of me, and unsure about whether my own feelings of guilt are justified. I might still have a nagging suspicion that I was in fact blameless, as manifest in spontaneous tendencies to assent to it, to view scolding me as unfair, to anger at those who resent my behavior, and to feel myself defective in feeling the guilt I do. While in my calmer moments I check these tendencies in line with a suspension of judgment, this nagging suspicion plays an evidential role in my deliberations about whether or not I am to blame, entering as data along with intuitions about others in similar circumstances and so on.

I think, however, that intuitively recalcitrant attitudes pose a problem for quasi-judgmentalism that is analogous to that posed by ordinarily recalcitrant attitudes to ordinary judgmentalism. The quasi-judgmentalist idea that valenced attitudes involve ethical perceptions entails that intuitively recalcitrant attitudes involve conflicting intuitions or sub-judgmental ethical evaluations. If I feel guilt for doing what I intuitively suspect to be blameless, my guilt for doing it must involve a conflicting intuitive appearance to the effect that I have indeed done something blameworthy. But it seems that I can feel guilt and suspect myself blameless without any such conflict in ethical intuition. If I were really conflicted in my evaluations or intuitions it looks as though I would have to manifest something like the following:

(I1) Conflicting inclinations to spontaneously assent to descriptions of my behavior as blameworthy or blameless, or at least use of both the view that I was to blame and the

view that I was not as starting points or data in inquiry into whether I was or not. This would involve conflicting tendencies to set burdens of proof in inquiry, or to be inclined to count the fact that a theory of blame predicts that my action is blameworthy both in favor of and against the theory.

(I2) Conflicting tendencies to spontaneously evaluate the ethical status and justifiability of related behavior. This would include conflicting tendencies to spontaneously view other lamp-breakers as culpable and blameless, to view those who anger at them as within and outside of their rights, and so on. Correspondingly, this would involve conflicting spontaneous tendencies to anger at and to refrain from angering at other lamp-breakers and those who criticize the lamp-breakers.

(I3) Conflicted tendencies to spontaneously view my feelings of guilt as appropriate and inappropriate. These would include conflicting tendencies to assent to descriptions of my guilt as appropriate, to feel good or bad about the guilt I feel on account of its perceived fittingness and unfittingness, and to be inclined to do something both to retain and to expiate my guilt on account of its perceived appropriateness or inappropriateness.

But it seems that I can feel guilt and intuit that I've done nothing blameworthy without manifesting any such conflicts. It seems that all of my tendencies could be in line with intuitions to the effect that I've done nothing blameworthy, and none in line with intuitions that my behavior was culpable. I could tend only to assent to descriptions of myself as blameless and tend unequivocally to set a burden of proof in favor of my blamelessness. I could unequivocally tend to spontaneous views of others like me as blameless and all who criticize us as out of line, with correspondingly univocal tendencies to anger. I could tend only to view my feelings of guilt as inappropriate, feel bad about them, and find myself inclined to expiate them on account of their inappropriateness, without any tendencies to the contrary. I could have all these tendencies, yet still feel guilt for breaking the lamp and have the suspicion that I was blameless. Surely, then, it seems that I feel guilt for doing something that I neither judge blameworthy nor intuitively perceive as such.

This pattern of argument must, of course, take care not to conflate the possibility that a certain tendency is very faint, or that it has been overridden by other, stronger tendencies, with the possibility that the tendency is completely absent.[66] Those of us who have measured the Müller-Lyer lines and know the story behind them may have no perceptible tendency to assent to descriptions of them as of unequal length or set a burden of proof in favor of the view that they are unequal. But, of course, we still perceive one line as longer than the other. It seems, however, that if we were genuinely unsure about their length (say we weren't sure if something was screwy about the lines or screwy about the ruler we were using to measure them), our perception of them as unequal would have some influence on our thinking. Indeed, if we were asked about the length of the lines without our having a chance to think (say in the context of a rapid-fire multiple choice quiz), I suspect that we might well have an inclination to judge one line longer than the other, which we would have to check in light of our background knowledge. But as we have seen, it looks as though one can feel guilt without manifesting any of the spontaneous tendencies associated with judging oneself blameworthy, even if one is unsure as to one's blameworthiness and even if one lacks time to think about it.

The quasi-judgmentalist might try claiming that the perception of blameworthiness involved in intuitively recalcitrant guilt is just very faint, and we mistake its weakness in the face of countervailing perceptions for its complete absence. But this looks *ad hoc*. If it looks possible to feel intuitively recalcitrant guilt without any of the symptoms of intuitively taking oneself to be blameworthy, it is a cost (e.g. an offense to parsimony) to have to say that the intuition is there but its symptoms are masked. Moreover, just as the judgmentalist could not use the bare presence of guilt as evidence for a judgment of blameworthiness, the quasi-judgmentalist cannot use the bare presence of guilt as evidence for an intuition of blameworthiness. We could of course call feelings of guilt "intuitions of blameworthiness" by courtesy, but then we would renounce all claims to the view that there is some theoretically important sense in which feelings of guilt involve "ethical content" about blameworthiness. Such a naming convention would at least mean that we could not object on account of our quasi-judgmentlaist-sounding claims that it is circular to use emotions like guilt in order to

[66] I am grateful to Neil Mehta for drawing my attention to this issue.

informatively analyze what we might call 'ethical content' in a stronger sense. This stronger sense of 'ethical content' could be the content of judgments and intuitions that play the roles in ethical inference, decision, and evaluation that we have been discussing. Ethical content in this (more meaningful) sense would be the content of those states of judgment and intuition that play roles in our normative thinking analogous to those played by empirical beliefs and judgments in our thinking about the external world.

We might do well to consider, however, if there could be versions of quasi-judgmentalism other than the view that valenced attitudes are states that involve construing things in terms of ethical concepts. It has been pointed out that even if we do generate perceptual images by deploying concepts, this etiology may not be part of the perceptual states themselves. As Tye (1995, 140-141) has argued, deploying the concepts VASE or FACES in viewing the Rubin vase may influence what we perceive by influencing how we decompose what we see into spatial parts. It seems that someone could decompose the features in the same way without deploying the concepts VASE or FACES – she might not even have these concepts.[67] To have perceptions of something as having a certain property one may not need to have the concept of that property.

A problem that arises for both judgmentlaism and the concept-deploying version of quasi-judgmentalism is that many beings seem to be capable of the relevant pro- and con-attitudes without possessing the corresponding ethical concepts. While some such attitudes, like guilt, may happen to be unique to psychologically typical humans above a certain age,[68] many others, like desire, anger, fear, etc., seem to be widely shared by non-human animals and very young humans. It seems quite likely, however, that at least some such animals and children have these attitudes without possessing ethical concepts like GOOD OUTCOME, OUTRAGEOUS ACTION, DANGEROUS SITUATION, and so on. The point is not that these beings do not speak languages – contra claims from certain quarters,[69] capacities for spoken language seem rather irrelevant to the ability to have concepts. Nor

[67] Tye does claim, however, that "one cannot see something as a rabbit…unless one has the concept rabbit. Likewise, it cannot appear to one that there is a rabbit by the hat unless one has the concepts rabbit and hat" (Tye 1995, 140). But I am not sure why this is so; if the perceptual contents are genuinely nonconceptual, it would seem that one could have the perceptions in question in virtue of the right kinds of image decompositions, nomic relations to rabbits and hats, and so on.
[68] See for instance Nunner-Winkler and Sodian (1988).
[69] Such as Davidson (1982).

do I think it plausible that the beings in question lack the ability to engage in the kind of domain general cognitive processes requisite for possessing concepts. I am actually rather confident that dogs and six-month-old humans have lots of concepts, just not the ethical concepts that would correspond to their valenced attitudes. This is because they seem to lack states that regulate their attitudes in the way our ethical judgments regulate ours, and they seem not to engage in the kind of response-guiding normative inquiry in which ethical judgments would play their inferential and evaluative roles.

It might, however, be open to the quasi-judgmentalist to claim that pro- and con-attitudes involve ethical content that is not conceptually embodied. It has been argued, for instance, that our phenomenal experiences and subdoxastic systems represent features of the world quite independently of our having concepts that represent these features. For instance, Tye (1995, 117) contends that "Hunger pangs represent contractions of the stomach walls when the stomach is empty," and Marr's (1980) theory of vision seems to attribute contents about such things as zero-crossings to certain representations in our visual systems. Beings of all kinds experience hunger and see things, but most of them (adult humans included) do not possess the concepts of a stomach contraction or a zero crossing. Perhaps the quasi-judgmentalist could try claiming that our valenced attitudes have ethical contents in the same way that states of hunger have contents about stomach contractions. The idea might be that feeling guilt for doing something represents your doing it as blameworthy in the same way that states of hunger represent stomach contractions. Hunter-gatherers in the late Pleistocene would have experienced hunger without any spontaneous tendencies to assent to corresponding views about stomach contractions or to set burdens of proof in favor of these views in a process of inquiry about such contractions. In the same way, feelings of guilt might involve representations of blameworthiness that are so cognitively impenetrable that they play no role in our thinking about blameworthiness. Similarly, just as our visual systems represent zero-crossings without our having the concept of a zero-crossing, the desires of non-human animals and young humans may represent the goodness of certain outcomes without their having to have the concept of a good outcome.

The first thing to note about this kind of quasi-judgmentalism is that it seems to renounce one of the central ideas that motivated judgmentalism in the first place - namely

the explanatory priority of ethical concepts to concepts of valenced attitudes and their fittingness. If the explanation of why guilt can be felt without judgments or intuitions of blameworthiness is that its representation of culpability is cognitively impenetrable, we would not seem to need the invoke the notion of blameworthiness in order to attribute guilt – any more, say, than hunger-gatherers in the late Pleistocene would have needed to invoke the notion of a stomach contraction in order to attribute hunger. This version of quasi-judgmentalism would seem unable to garner motivation from our descriptions of the phenomenology of guilt as "feeling like you've done something blameworthy." If the representation of culpability is cognitively impenetrable, why would we be expected to describe its feeling as such? Pleistocene hunger-gatherers wouldn't have described hunger as "feeling like your stomach is contracting." The proponent of the cognitive impenetrability version of quasi-judgmentalism will presumably need to explain the description of guilt's phenomenology in a different way. Perhaps we usually feel guilt when and only when we think that we're to blame, and we are describing guilt's phenomenology by reference to how we typically feel in such a situation. But if this is right, then our descriptions of guilt's phenomenology don't really support the conceptual priority of the notion of blameworthiness over the notion of guilt after all.

A second problem for this version of quasi-judgmentalism is that it is very hard to see how pro- and con-attitudes could come to have cognitively impenetrable ethical content. Our evidence for thinking that hunger represents stomach contractions and that visual states represent zero-crossings consists in the fact that these mental states bear certain nomic relations to stomach contractions and zero-crossings. Moreover, our best theories of content suggest that some such nomic relations are what make it the case that these mental states represent these states of the world.[70] Our pro- and con-attitudes may well bear similar nomic relations to various states of the world. Distally, our desires for a state of affairs may lawfully co-vary with things like the fact that it will conduce to the pleasure, bodily integrity, or status of our selves or those we care about. Proximally, our emotions may co-vary with the fact that our bodies are undergoing the kinds of changes that constitute and accompany the arousal of the sympathetic nervous system.[71] But why

[70] See for instance Stampe (1977), Dretske (1981, 1988), and Fodor (1987, 1990).
[71] See for instance D'Amasio (1994) and Tye (1995).

should we think that our desires or emotions bear nomic relations to ethical facts? As we saw in chapter 1, unreduced ethical facts look superfluous from the standpoint of explaining the etiology of our attitudes. Ditto for alleged synthetic identities between ethical facts and other facts, like those about what will be pleasing to whom. The only option left would be that our valenced attitudes are nomically related to certain facts that we can describe in non-ethical terms but that analytically entail certain ethical facts. But it looks difficult to see how anything ethical is *analytically entailed* by the facts to which our valenced attitudes (and those of infants and non-human animals) are nomically related.

Yet the most devastating problem for the cognitive-impenetrability version of quasi-judgmentalism is probably that it repudiates the judgmentalist's explanation of how our valenced attitudes are responsive to our ethical judgments, and offers nothing to put in its place. As we saw in section 2.1, a striking feature of our pro- and con-attitudes is that we can evaluate them, not only as states that we should get ourselves to have, but as states that it is fitting or appropriate to have. Just as judgments that a belief is epistemically warranted have a direct propensity to cause us to have it, these judgments that a valenced attitude is fitting or warranted have a similar propensity to cause us to have it without our having to do anything to bring it about that we do. What the judgmentalist offered us was an attractive picture of the causal powers of fittingness judgments in terms of the causal powers of judgments about warrant for belief. If your assessment of the evidence in favor of an ethical judgment can directly cause you to have it, and you only count as having a valenced attitude if you have the right ethical judgments, your evidential assessments of ethical judgments will directly control what valenced attitudes you have. Assessments of an attitude's fittingness were thus explained as evidential assessments of the ethical judgment it involves, with the direct power to retain or destroy the judgment and thus the attitude that involves it.

While the concept-deploying version of quasi-judgmentalism was able to make more room for the resistance of our valenced attitudes to control by our ethical judgments, it was able to retain the same basic picture of how our valenced attitudes responded to our ethical judgments when everything was working properly. We can judge something to be one way but construe it to be another, which might well be what

happens when we see the Müller-Lyer lines as unequal in length but judge their lengths to be the same. But what we believe or judge about something does seem to have a certain propensity to influence how we construe it, which operates directly and without the need of our intentional control. Thus, if you believe that the duck-rabbit is a picture of a rabbit or that the Rubin's vase is a picture of a vase (because someone has told you that this is what the artist was drawing and you are not wise to the fact that they are ambiguous images), you will be likely to construe them as such and have perceptual experiences of a rabbit or a vase. In the same way, the proponent of concept-deploying quasi-judgmentalism can hold that the ethical construals involved in our pro- and con-attitudes are directly influenced by our ethical judgments, which are in turn directly influenced by assessments of the evidence in favor of them.

But the quasi-judgmentalist who holds that pro- and con-attitudes involve cognitively impenetrable ethical evaluations has absolutely no explanation of how our ethical judgments (and intuitions) would influence these attitudes. That they cannot do so is largely just what it is for these evaluations to be cognitively impenetrable. Believe what you will about whether your stomach is contracting – it will have no effect on whether you experience hunger pangs. As Tye (1995, 118) reports, a patient who was well aware that his stomach had been removed still experienced entirely normal feelings of hunger. Similarly, it is often held that feelings of pain represent tissue distortion or damage at the place where the pain is felt to be. But one of the pesky things about phantom-limb pain is that knowing there to be no tissue damage or distortion at the felt location of the pain seems to have no propensity to take the pain away.[72] Any cognitively impenetrable ethical content possessed by our emotions and desires would be similarly impervious to influence by our ethical judgments.

The quasi-judgmentalist might try to accept this break with traditional judgmentalism and seek to account for the correlation between ethical judgments and valenced attitudes in terms other than the former causing the latter. She might try for a kind of "preestablished harmony" explanation, according to which we have ethical judgments corresponding to our valenced attitudes (when we do) because the cognitively impenetrable representations embodied in our attitudes and the cognitively penetrable

[72] See Tye (1995, 112).

representations constituted by our ethical judgments have been triggered by the same features. But this would fail to account for the fact that different ethical outlooks tend to give rise to different valenced attitudes even in the presence of the same merely descriptive beliefs and perceptions. Of people who hold the same merely descriptive beliefs, those who think it obligatory to keep a certain promise will tend to feel more guilt for breaking it than people who think it permissible to break it, people who think it obligatory to do a lot to the poor will tend to feel more guilty for living high and letting die than people who think that they don't owe the poor anything, people who think it intrinsically good for killers to be killed in turn will tend to have a stronger desire for the death penalty to be on the books than people who think this kind of retribution isn't itself good – and so on for just about every ethical judgment and corresponding valenced attitude. These generalizations support non-backtracking counterfactual conditionals – if you were to keep your descriptive beliefs fixed but change your mind this way or that about such and so ethical issue, you would tend to have the valenced attitudes that correspond to your new ethical judgments. Of course the tendency is not iron clad, and as we have seen, changes in ethical view can take time to overcome deeply entrenched patterns of attitude. But the tendency is still there, and more often than not the influence is immediate. Changing your mind about whether you were obligated to keep a promise, return some property, or give someone some help tends immediately to alter whether you feel guilt for your failure to do so. Changing your mind about whether it's better for a few very poor people to be lifted out of poverty or many not so poor people to be made a bit better off tends immediately to alter whether you prefer a policy that will bring about the former outcome to one that will bring about the latter.

2.5. The Governance of Affect, Attention, and Motivation

We thus seem to need some story about how ethical judgments have a propensity to cause us to have the pro- and con-attitudes the fittingness of which they entail. Judgmentalism and the concept-deploying version of quasi-judgmentalism seemed to offer an attractive story about how this works, but they got into trouble with recalcitrant attitudes. The

cognitive-impenetrability version of quasi-judgmentalism had a way of getting around the recalcitrant attitude problems, but it had no story about how ethical judgments govern pro- and con-attitudes. This is all very bad for the view that valenced attitudes involve ethical content. But let us return to the judgmentalist story about how ethical judgments guide valenced attitudes. I believe that closer inspection reveals a critical lacuna in this story, and that understanding this is of central importance to understanding the relationship between ethical judgments and valenced attitudes.

As we saw in section 2.2, we can make bloodless ethical judgments – we can judge ourselves culpable without feeling guilt, judge a state good without desiring it, and so on. This is because attitudes like emotions and desires involve phenomenal, motivational, and attention-directing elements that ethical judgments can lack. As such, we cannot equate ethical judgments with pro- and con-attitudes. Rather, it looked most plausible for the judgmentalist to say that having an ethical judgment is part - but not all - of what it is to have a valenced attitude. The other parts are presumably the phenomenology, motivation, and attention-direction that bloodless ethical judgments lack.

According to the judgmentalist's story, your ethical judgments determine which pro- and con-attitudes you have solely by regulating the ethical representations that are essential parts of these attitudes. If you feel guilt for what you have done but come to think you are blameless, then whatever you might still feel it can no longer count as guilt. Similarly, if you come to judge that a certain outcome would be good, and you happen to have the right phenomenology, motivation, and focused attention on the scene, you can count as desiring that state where you did not count as desiring it before. But what we lack on this picture is any account of how changes in ethical representation can affect the presence or absence of those features of emotion and desire that are not themselves ethical representations.

The phenomenon of bloodless ethical judgments shows that you can make an ethical judgment without having the kind of phenomenology involved in having the corresponding valenced attitude. But it is still open to the judgmentalist to insist that you cannot have the exact kind of subjective experience characteristic of a particular valenced attitude unless you make the corresponding ethical judgment. Thus, you might need to

add some kind of phenomenal element to judgments of blameworthiness to get feelings of guilt, but you would not feel exactly as you do when you feel guilt if you kept the added phenomenal element but took away the judgment about culpability. For clarity, I shall refer to the phenomenal element that the judgmentalist thinks you need to add to an ethical judgment to get the corresponding valenced attitude as 'the affect' involved in that attitude.[73]

What the judgmentalist holds, then, is that if you happen to have the right kind of affect, motivation, and directed attention, adding the corresponding ethical judgment makes the resulting complex count as a certain emotion or desire, and taking away the judgment makes it cease to count as such. If you start out with the right kind of negative feeling, amends-like motivation, focus on what you have done, etc., then if you judge that you have done something blameworthy you get to count as feeling guilt, and if you stop judging that you have culpably offended you don't get to count as feeling guilt. But there is no mechanism by which the presence or absence of the judgment has any effect on whether the affect, motivation, and focused attention are themselves present or absent. The role of the judgment is simply (as it were) to "brand" any existing affect, motivation, and focused attention as "guilt," and the role of ceasing so to judge is simply to remove the "guilt brand" from these non-judgmental or *conative elements* without in any way altering their existence. In the same way, if you don't start out with the prerequisite affect, motivation, and focused attention, any judgments of yourself as having culpably offended will simply be bloodless. The judgmentalist story lacks a mechanism by which ethical judgments could themselves bring about the kind of conative elements that she agrees we need to have in order to count as having valenced attitudes.

The inability to explain how ethical judgments influence the conative elements involved in having pro- and con-attitudes is, I think, a very big problem for the judgmentalist's story. As we have seen, judging a valenced attitude fitting or unfitting is distinctive in that it has a direct propensity to cause one to have it or refrain from having it. But the propensity of fittingness assessments to alter our valenced attitudes seems to be a propensity to alter all parts of our valenced attitudes, conative elements included.

[73] This is, I believe, rather close to what is conventionally called 'affect', though I do intend to use the term stipulatively to refer to the phenomenal element that can (all parties can agree) be had without the making the corresponding ethical judgment.

When everything is working smoothly, coming to judge that you've done nothing wrong causes you to stop feeling guilty, not just by re-labeling the affect, motivations, and preoccupations you have, but by causing you not to have them at all. Suppose that you assign a student a low grade, and at first you think this happened in part because you owed him a better explanation of the assignment. You feel guilty for what happened, with the characteristic negative affect, motivation to make things up to the student, dwelling on how you might have done more to explain the assignment, and so on. But suppose that, in reflecting upon similar cases and how much in general you think instructors should be expected to explain, you come to judge that you did not have an obligation to explain the assignment better than you did. In many if not most such situations, you will not simply retain bad feelings, motivations to make things up to the student, and focus on what more you could have done, while agreeing that you need to re-label them something other than guilt. Rather, your coming to judge that you did nothing to blame will tend to directly remove the bad affect, motivations to make amends, and dwelling on the past.

In the same way, coming to judge that you have done something culpably wrong seems capable of causing guilt, not simply by labeling conative elements that already happen to be present, but by bringing new affect, motivation, and focused attention into existence. Suppose, for instance, that Bill finds a wallet and pockets the money without hesitation, either without thought for the person whose wallet it is or confident in the view that anyone careless enough to drop his wallet does not deserve it to be returned. One day some time later Bill is sitting in a café, and notices that the man next to him has his wallet on the table and is distracted with something. It occurs to Bill that the man is so distracted that he could easily take his wallet without his noticing, but it also occurs to him that it would be wrong to do so. But, Bill thinks, there really doesn't seem to be anything so relevantly different between pocketing a careless person's money when he is present and pocketing his money once he has left. As a result he concludes that he was culpably wrong to have pocketed the money that he found, and feels guilty for doing so. In many such cases, it seems that this would cause Bill to have brand-new negative affect towards his past action, motivation to find try to find the person whose wallet he took and restore the money to him, and tendencies to dwell on what he did and how he so easily

could have done otherwise. It was not as though Bill somehow had this affect, motivation, and attentional focus before he came to judge his act blameworthy, and that the judgment merely re-labeled it 'guilt'. The judgment of culpability itself seems to have caused such conative elements where there were none before.

Much like the proponent of the cognitive-impenetrability version of quasi-judgmentalism, the judgmentalist might here try to explain away such apparent causation as correlation engendered by a hidden variable. The judgmentalist might contend that our ethical judgments tend to be correlated with the conative elements of our valenced attitudes because the same kinds of features tend to trigger both our ethical judgments and our corresponding affect, motivation, and attentional focus. Perhaps, for instance, the perception that one's action pained someone, or interfered with his projects, or broke a promise, or whatever both (i) causes one to feel the conative elements of guilt towards one's action and (ii) causes one to judge one's action blameworthy. But the problem here (much like the problem for the aforementioned quasi-judgmentalist) is that such an account cannot explain how different ethical outlooks tend to give rise to different conative states in the presence of the very same descriptive beliefs and perceptions. Indeed, the examples just given seem to illustrate this phenomenon. When you changed your mind about whether you wronged the student, it was not as though you came to represent you action as having a new (merely) descriptive feature – you just thought through cases and principles about what instructors seem to owe students, and came to think that your behavior was relevantly unlike instances of wrongful behavior. Yet this change in ethical outlook seemed likely enough to be sufficient to cause you to have the affect, motivation, and attention characteristic of guilt. In the same way, Bill's change of mind about the moral status of his pocketing the lost money was in no way caused by changes in his merely descriptive beliefs about what his action was like. What changed his mind were simply his intuitions that it would be wrong to pocket the money on the table and that there was no very significant moral difference between pocketing that money and pocketing the lost money.

Similar points could be made about the cases we discussed in section 2.4. As we saw, if Bobby and Suzy have the same merely descriptive beliefs about living high and

letting die, but Bobby thinks it deeply wrong to do so and Suzy does not, Bobby will tend to feel more guilt for doing so than Suzy will. But it is not as though Bobby and Suzy have the same negative affect, motivations to make it up to the poor, and focus on what they have done, where Bobby's can be labeled "guilt" and Suzy's can't. It is rather that Bobby will tend to feel negative affect, be motivated to make amends, and focus on his behavior where Suzy simply will not. The tendency is by no means iron clad, but coming to judge like Bobby tends to generate the conations characteristic of guilt, while coming to judge like Suzy tends to destroy or prevent them.

The judgmentalist could, of course, add to her story, claiming that we just happen to have certain tendencies to respond with affect, motivation, and attentional direction to our ethical assessments. For instance, to explain the sensitivity of our motivations to our ethical judgments, she could, for instance, claim that we have a standing *de dicto* desire to make amends for doing what is blameworthy and a standing *de dicto* motive to bring about whatever states are good. One problem about this kind of claim, as we shall see later on, is that it seems to paint us as having what Bernard Williams would call "one thought too many" – rather than being motivated to bring about what is good (say children's' happiness) as an end in itself, we come out looking like we want to do it only as a way of making it the case that good states of affairs obtain. Another problem with the posited of tendencies to respond to ethical judgments with conation is that until more is said it is unclear how they can square with the phenomenon of bloodless judgments. The very same people sometimes respond with conation to the judgment that they have culpably offended and sometimes do not. What kind of tendency would explain this? A standing desire to make amends for whatever is blameworthy would seem to be a poor explanation of why we sometimes lack a desire to do so.

But the main thing I want to emphasize for now is that the addition of such posited tendencies to respond to ethical judgments with conation seems to detract from the explanatory virtues of the judgmentalist's account of how our ethical judgments are supposed to guide our valenced attitudes. These seem somewhat *ad hoc*, decrease the account's explanatory power, and add an unpalatable layer of complexity to it. Initially, it looked like the judgmentalist was going to reduce the attitude-guiding propensity of our fittingness assessments to the attitude-guiding propensity of our evidential assessments.

The similarity between the influence of judging guilt fitting on feeling guilt and that of judging a belief justified on having the belief was supposed to be explained by the fact that the former was an instance of the latter. Judging guilt fitting, the story went, is controlled by an evidential assessment of blameworthiness, which directly controls the blameworthiness judgments involved in guilt and hence the guilt itself. But now we have seen that this cannot be right – the judgment that one has done something blameworthy or that guilt is fitting has a direct effect on the conative elements of guilt as well. To explain this, the judgmentalist is forced to posit additional, unexplained tendencies of conation to respond to ethical judgment. Wouldn't it at least be nicer if we had an explanation of how all the parts of our valenced attitudes respond to our fittingness assessments together, as a corporate body?

2.6. What Are Ethical Judgments Anyway?

Judgmentalists (and quasi-judgmentalists) are committed to the idea that we can understand ethical judgments (or percepts or representations) quite independently of the pro- and con-attitudes that they claim involve such judgments. But can we? What exactly is it to judge that an act is blameworthy or that a state is good? One thing that is striking about ethical concepts is the diversity of things that people can coherently (if in many cases quite falsely) judge to fall under them. Consider, for instance, the wide diversity of things that people have coherently held to be blameworthy. Some of these things would include inflicting harms upon others, failing to prevent harms to others, defecting in the presence of collective action problems, and failing to respect the autonomy of other agents. But they would also include all manner of apparently miscellaneous behavior, including sexual practices, drug use, violations of etiquette, "playing God" by engaging in cloning or genetic modification of organisms (quite apart from its effects on any being's welfare), failures to adhere to certain religious practices, stringing together certain phonemes (in the form of curse words[74]), and so on. It should be emphasized that these kinds of apparent miscellany can and have been coherently

[74] I am grateful to John Ku for this example and help with this list generally.

thought to be intrinsically wrong and blameworthy quite apart from beliefs about their contribution to anyone's welfare or autonomy.

Similarly, consider the dizzying array of states of affairs that people have coherently held to be intrinsically good. These would include states in which sentient beings live long and happy lives free from pain, unmerited advantages are equally distributed, people have knowledge, people achieve things, natural beauty is preserved, and artistic endeavors flourish. But they would also include all kinds of apparently miscellaneous states of affairs, including those in which racial purity is maintained, people are chaste, traditions are followed, deities and ancestor spirits are obeyed, women know their place, and so on. Let me again emphasize that people can (and have) coherently thought the latter kinds of states intrinsically good, and good apart from their further effects on anyone's happiness or anything else.

If this is right, then it does not look as though we are going to be able to analyze ethical judgments in some simple kind of descriptive terms. An attempt to analyze the notion of a blameworthy act as an act that fails to maximize happiness, or is hated by deities, or violates autonomy, or what have you would conflict with our intuitions to the effect that someone could think an act had none of those features and yet was blameworthy all the same. Ditto for attempts to analyze good states of affairs as states that maximize pleasure, or maximize a given distribution of an objective list of something, or what have you. Some such view about which acts are blameworthy or which states are good might very well be *true*; the coherence of denying the view simply shows is cannot be a conceptual truth. It takes substantive work to figure out which acts are intrinsically blameworthy and which states are intrinsically good, and when people disagree about this their disagreement is not about which view is coherent.

If we can't analyze ethical judgments as judgments about a simple kind of descriptive fact that we can describe in non-ethical terms, how can we understand them? We could try saying that they are absolutely primitive and such that we can say nothing informative about them. This might seem to be a rather unhappy place on which to rest our understanding of the emotions and desires that the judgmentalist insists we must explicate in ethical terms. For one thing that it *does* seem informative to say about ethical judgments is that they have a direct propensity to shape our valenced attitudes –

or, if you like, the affect, motivation, and directed attention that those attitudes involve. This direct influence on our valenced attitudes is just like the influence of our judgments about our epistemic reasons on our beliefs. Another thing that it seems informative to say about ethical judgments is that our basic ethical thought proceeds by means of reflective equilibrium methods that seek out a best unification of our (non-debunked) ethical intuitions of all levels of generality. In this respect basic ethical inquiry is just like basic inquiry into what to believe or what to do. As with our intuitions about what to believe and do, the denial of the ethical intuitions we rely upon in our reflective equilibrium methods seems coherent; it just seems false.

Since these are some of the most informative things we can say about what ethical judgments are, we need to consider reversing the judgmentalist order of explanation. Suppose we tried taking as basic the state of judging that an attitude is warranted or justified. The kinds of attitudes we can assess for warrant include both our beliefs and our valenced attitudes, where we assume that our valenced attitudes *do not* involve ethical judgments. Our valenced attitudes are to be understood, rather, as something like the syndromes of directed affect, motivation, and attentional focus that we saw our ethical judgments can lack. On this picture, we can directly assess the warrant of both of our two main kinds of domain general mental states: our representational or "cognitive" states, and our motivational or conative states. Judging a belief warranted is the same thing as judging that we have epistemic reason to have it, and judging a valenced attitude warranted is the same thing as judging that it is fitting. A distinguishing feature of judging an attitude warranted is its propensity to directly cause you to have that attitude. On this picture, we do not seek to reduce the influence of fittingness assessments on valenced attitudes to that of epistemic assessments on beliefs. Rather, we explain both as equally basic instances of the sensitivity of attitudes to assessments of their warrant. The picture of attitude regulation that we get by taking fittingness assessments as a basic kind of warrant assessment is illustrated and contrasted with the judgmentalist picture in figure 3 below.

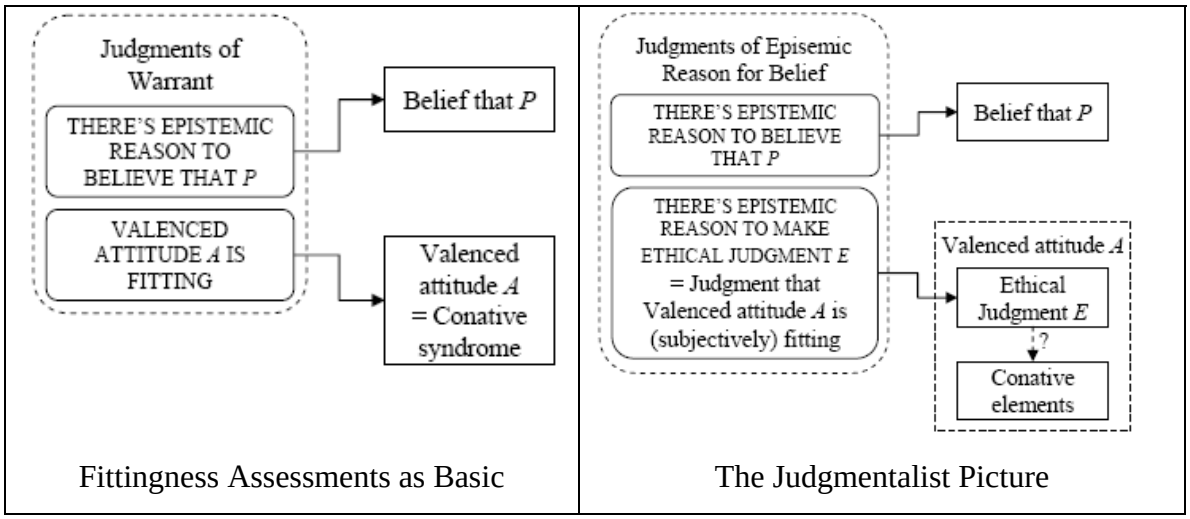

Figure 3: Attitude Guidance on the Picture of Fittingness Assessments as Basic vs. The Judgmentalist Picture of Attitude Guidance

We might in this way understand judgments that a valenced attitude is fitting as a basic kind of mental state that guides the attitude, enters into inferential roles with other states of its kind, and responds to reflective equilibrium methods of basic normative inquiry. Note that we seem to be able to understand these features of fittingness assessments without our having to deploy our ethical concepts. As such, we can actually use fittingness assessments understood in this way to give informative analyses of different ethical concepts as concepts of the fittingness of different pro- and con-attitudes. In particular, each ethical judgment is to be analyzed as the fittingness judgment it entails: ethical judgments entail judgments that certain attitudes are fitting because they are identical to these fittingness judgments. Examples of these analyses would include the following analyses of our concepts of MORAL BLAMEWORTHINESS and GOOD (or BETTER) STATES OF AFFAIRS:

> **Fitting Attitude Analysis of Moral Blameworthiness**:
> To judge that someone's act is morally blameworthy is to judge that it is fitting for him to feel guilty for having done it, and fitting for others to be angry at him for doing it.[75]

[75] This analysis is given by Gibbard (1990, 40-45, 126-127), who cites similar analyses by Ewing (1939) and Mill (1863) as important precedents.

Fitting Attitude Analysis of Good (and Better) States of Affairs:

To judge that state of affairs *S* is good is to judge that it is fitting for us to have a pro-attitude towards *S* (e.g. desire, wish, or be glad that *S* obtains). To judge that state of affairs *S* is better than state of affairs *S*' is to judge that it is fitting for us to prefer *S* to *S*'.[76]

These fitting attitude [FA] analyses offer us a way to explain the normative features that ethical judgments share with each other and with judgments about epistemic reasons for belief. First, they offer us a straightforward explanation of how ethical judgments have the same kind of direct influence on our conations that judgments about epistemic reasons have on our beliefs. We are understanding valenced attitudes like guilt and desire as conative syndromes, or patterns of affect, motivation, and attention that are felt and directed towards things like individuals, actions, and states of affairs. Judgments that the attitudes comprised by these syndromes are warranted exert causal pressure on our tendencies to have the affect, motivation, and attentional focus that comprise them. Thus, judgments of blameworthiness influence the conations that comprise guilt and anger and judgments of the value of states of affairs influence the conations involved in desires and preferences because these judgments are judgments of warrant for these conative syndromes. The FA analyses of ethical concepts, together with the picture of fittingness judgments as basic, subsume the causal influence that we see exerted by each ethical judgment on different valenced attitudes under the general phenomenon of the governance of attitudes by judgments of their warrant. In this way it gains us a unified explanation of the influence of these judgments, which is also unified with our understanding of the influence that judgments about epistemic reasons have on beliefs.

While the FA analyses can in this way explain the sensitivity of our valenced attitudes to our ethical judgments, they can readily explain how our ethical judgments can be bloodless and how our valenced attitudes can be recalcitrant. Judging an attitude warranted or unwarranted has a propensity to cause you to have it or refrain from having it, but the tendency competes with other causal mechanisms and is by no means iron clad. As we saw in Chapter 1, you can judge that you have overwhelming epistemic reason to

[76] See e.g. Ewing (1939), Gibbard (1990), and Scanlon (1998)

believe the many worlds interpretation of quantum mechanics, but find yourself unable to do so. The influence exerted on our beliefs by our epistemic assessments can in this way be insufficient to overcome the inertial resistance of beliefs to change in such radical ways in the absence of certain vivid cues. In the same way, a person who grew up with beliefs in magic might conclude that the epistemic case against believing in such things as monsters, goblins, and deities is overwhelming. But he might still find himself avoiding certain places or actions in ways that seem best explained as the result of his having a higher credence in the existence of mythical beings than he thinks he ought to have. In such cases the judgment that a belief is unjustified is insufficient to expiate it in the face of well-worn habits of thought. The FA analyst can explain our bloodless ethical judgments and recalcitrant attitudes as analogous failures of our valenced attitudes to respond to assessments of their warrant.

The FA analyses also offer us a way to subsume the relationship between ethical intuitions and ethical thought under the more general relationship between intuitions of warrant and thought about warrant. The proponent of FA analyses will treat ethical intuitions as intuitions to the effect that certain valenced attitudes are fitting or warranted. Since she holds that pro- and con-attitudes involve no intuitive or sub-judgmental content about ethics or fittingness, she has a ready explanation of how valenced attitudes can be intuitively recalcitrant without automatically entailing conflicts in intuition. Like judgments of warrant, intuitions about warrant have a propensity to guide our attitudes, but the propensity is not always decisive. More generally, intuitions of warrant play the kinds of roles we saw in section 2.3. In addition to guiding attitudes in the heat of the moment, intuitions of warrant are less responsive to reasoning than judgments of warrant, and they play a role as starting points to basic normative inquiry into which attitudes actually are warranted. As we saw from our example in Chapter 1 of the intuition that we should believe theories that posit fewer things, the role of intuitions of warrant as starting points involves the debunking of some of these intuitions by theories that can best explain the totality of them. The account of ethical intuitions as intuitions of warrant unifies our understanding of the role they have in ethical inquiry with our understanding of the role epistemic intuitions have in inquiry into what to believe in light of our evidence. The use of reflective equilibrium methods in substantive ethical inquiry and their use in

substantive inquiry about what to believe are both instances of the fact that we conduct our basic inquiry into which attitudes are warranted by seeking out a best unification of our non-debunked intuitions of warrant.

Finally, it seems that FA analyses of ethical concepts give us a way to explain what is common to the content of ethical judgments, no matter how diverse the range of coherent such judgments may be. As we saw, we seem to be able to understand judgments that all sorts of acts are intrinsically blameworthy – not just those of inflicting harms and violating autonomy, but those of engaging in certain sexual practices and uttering curse words. What seems to explain our ability to understand judgments that such disparate acts are blameworthy is our taking it that the people who make them think that the action in question warrants guilt on the part of the person who performs it and justifies outrage or indignation on the part of others. And what seems to explain our ability to understand these judgments of warrant is our taking it that those who make them have the tendencies towards attitude, inference, and reflective-equilibrium thinking that would be characteristic of such warrant assessments.

Anyone who thinks that it is culpably wrong to curse will have a propensity to feel guilt for doing it and get angry or indignant with people who do. She will, that is, have a propensity to have the characteristic affect, motivations, and attentional focus of guilt towards her own cursing, and to have the characteristic conations of anger towards the cursing of others. Should such a person curse and fail to feel guilty, she will tend to regard her lack of affect as a manifestation of something going wrong with her. She will tend to infer things from her judgment, like that others are within their rights to berate her if she curses and that moral theories that make no room for the intrinsic culpability of cursing are false. If her judgment is supported by direct intuition she will tend to defend intuitions about the intrinsic culpability of cursing from attempted debunking arguments, and if her judgment is inferred she will tend to look for explanations of how the intrinsic blameworthiness of cursing is entailed by a theory that makes best sense of her other normative intuitions.

Similar remarks could be made for our ability to understand bizarre views about which outcomes are good. Anyone who thinks that racial purity is intrinsically good will think that desires or preferences for racial purity are warranted. Thinking this will have a

propensity to cause one to prefer racial purity – to have positive affect towards states of racial purity, to be motivated to bring them about, and to direct attention to opportunities to create or preserve racial purity. When one thinks these preferences warranted but one's propensity to have them fails to cause these conations towards racial purity, one will regard one's response as somehow deficient, and look for ways of fixing it. The judgment that preferences for racial purity are warranted will tend to give rise to related inferences about what is good and what is to be done, as well as attempts to either defend intuitions that desires for racial purity are warranted or show that their warrant would best explain our other normative intuitions.

Foot (1959) famously suggested that one could not coherently think that hand-clasping is intrinsically good. I must confess that I cannot see any reason why there would be any more of a problem with the coherence of this view than the view that cursing is intrinsically blameworthy or that racial purity is good in itself. But however this may be, the thought seems perfectly coherent (though of course perfectly false) as the view that intrinsic desires for hand-clasping are warranted, with the attendant tendencies towards such desires, inference, and reflective-equilibrium thinking. Is there anything wrong with this way of understanding judgments about what is good? Part of the answer will have to wait until the next chapter, in which we will discuss how exactly ethical judgments are related to judgments about what to do. But for now we can observe that the understanding of judgments about goodness and other ethical judgments provided by FA analyses seems to capture the central causal, inferential, and epistemic features of these judgments. Armed with this, the fact that these analyses give us a way to interpret judgments about the intrinsic goodness of hand clasping seems to be so much the worse for Foot's more stingy views about what is coherent.

FA analyses of ethical concepts and the view of fittingness assessments as basic fly in the face of the two claims about valenced attitudes that, as we have seen, are taken to support judgmentalism (and can be just as easily extended to support certain versions of quasi-judgmentalism). The first is that we cannot distinguish our pro- and con-attitudes from one another without saying that they involve different ethical evaluations. In our discussion of the cognitive-impenetrability version of quasi-judgmentalism, we

already saw one line of response to this claim. Pro- and con-attitudes involve a distinctive phenomenology, and phenomenological differences are notoriously difficult to describe in spoken language. This offers us a way to debunk the claim that valenced attitudes must involve ethical evaluations on the grounds that we say things like "guilt feels like you've done something awful" and "when you desire *S* it feels like *S* would be good." Given the propensity of valenced attitudes to respond to judgments of their warrant, we will *usually* or *often* feel guilt for doing something just in case we think we are blameworthy, and we will *typically* desire a state of affairs when and only when we judge it good. Asked to describe these attitudes, we may simply be trying to communicate something informative by referring to the way one tends to feel in situations that typically engender them.[77]

But the fact that we sometimes use ethical language to describe the phenomenology of our valenced attitudes might not be the only reason the judgmentalist thinks we need constituent ethical evaluations to individuate valenced attitudes. Perhaps she just can't see how anything else *could* individuate our valenced attitudes. Arguments in favor of (quasi-) judgmentalism have sometimes held that without constituent ethical evaluations the only thing we have left is a kind of undifferentiated positive or negative affect, with nothing to tell the difference between, say, guilt and shame.[78] But from what we have seen so far this is simply false. First, the conations that comprise emotions and desires involve distinctive patterns of motivation and directed attention in addition to affect. Guilt, as we have seen involves motivations to do things that we can describe as "making amends" – restoring the world to its prior state, catering to offended parties, and treating oneself in "self-punishing" ways that would ordinarily look aversive. Shame, as we will see, involves a different set of motivations – mainly to develop or exercise one's competence.

---

[77] Interestingly enough, there seem to be other occasions on which it can be informative to use the names of valenced attitudes to describe the phenomenology that typically accompanies ethical judgments, as in:

Q: "How do you usually feel if you think someone's culpably wronged you?"
A: "Angry"

Q: "How do you usually feel about something if you think it's good?"
A: "You want it, wish for it, or are glad of it"

[78] See Solomon (1976).

Second, there really does seem to be a phenomenal difference between different pro- attitudes and between different con-attitudes. Anyone who claims that guilt and shame feel the exact same seems either to be using the terms 'guilt' and 'shame' to describe different mental states than the rest of us, or to be misled about his own subjective experiences. The judgmentalist might claim that this difference is a consequence of different ethical evaluations, but at this point the burden seems to be on him to show that this is the case. The phenomena of recalcitrant and intuitively recalcitrant attitudes combined with the failures of the judgmentalist picture to explain the guidance of conation seem to count seriously against her story of what individuates phenomenology. We have as yet no direct reason why those who deny that valenced attitudes involve ethical evaluations cannot point to different phenomenal characters to individuate valenced attitudes.

Third, the affect involved in different valenced attitudes can take different kinds of entities as intentional objects. Desiring a state of affairs in the directed attention sense involves having positive affect towards that state of affairs, while morally esteeming someone involves having positive affected directed towards that person. Finally, while this is a point we shall have to return to later, valenced attitudes have relationships with one another that can help to distinguish them. Guilt is related to anger or outrage in a way shame is not, and shame is related to contempt or scorn in a way guilt is not. There is a distinctive tendency to feel anger at other people for doing what you tend to feel guilt for doing yourself, and a distinctive tendency to have scorn or contempt for people who do what you feel shame for doing yourself. These relationships between valenced attitudes, together with distinctive phenomenology, objects of direction, motivation, and patterns of directed attention provide a rich set of resources with which to individuate different attitudes. Ditching the idea of constituent ethical evaluations leaves us with far more than undifferentiated affect.

The second claim about valenced attitudes taken to support judgmentalism as against FA analyses and the view of fittingness assessments as basic is the claim that it is conceptually impossible to have valenced attitudes towards certain kinds of objects. Of course, the proponent of FA analyses might allow that this is true to a certain extent; it might be part of what guilt is that it must be felt towards one's own acts or omissions,

and thus that it is conceptually impossible to feel guilt about the number two or the person Julius Caesar. But the FA analyst seems to have as good an explanation of this sort of thing as the judgmentalist – the former explains this in terms of the possible objects of the conations involved in guilt while the latter explains this in terms of the possible objects of culpability assessments. Both will insist that his explanation is more fundamental, but the determination of who is correct seems to devolve entirely to the impendent virtues and vices of their theories.

What might seem to count in favor of the judgmentalist's position would be examples of things that are of the right general ontological type but still ruled out as objects of certain valenced attitudes as a matter of conceptual necessity. Foot, as we have seen, claims that one cannot take pride in something unless one sees it as "in some way one's own," contending that one could not take pride in the sky or sea unless he were under a delusion that he kept the sky from falling or the sea from drying up. But, as D'Arms and Jacobson (2003) have argued, the only sense in which you must see something as "your own" in order to take pride in it is a completely trivial one in which to see something as relevantly "your own" is nothing more than to be able to take pride in it. Foot seems to be wrong that a person would need to have extravagant beliefs about his causal role in the maintenance of the sky and sea in order to take pride in them. We know, for instance, that people often take pride in their local parks and neighborhoods even though they know very well that they have no causal role in maintaining them. Suppose we came into contact with space aliens and had regular social relations with them. It seems that someone showing visiting aliens from another planet around earth might very well take pride in the sky and sea as he shows them off in the exact same way a tour guide from Chicago might take pride in the Sears Tower.

Similar remarks might seem to apply to Kenny's claim that it is impossible to feel remorse "for something in which one believes one had no part." Kenny seems to suggest that in order to feel remorse for crop failures in Vietnam one must hold extravagant beliefs like those to the effect that the crops failed because one's prayers were inadequate. But in the phenomenon of survivor guilt, people seem to feel remorse for the fact that others were harmed instead of them. It seems possible for them to feel this remorse while fully believing that who was killed was beyond their control. If there is

any sense in which one must believe one "had a part" in something in order to feel remorse for it, it may simply be the trivial sense in which to believe one "had a part in something" is just to be able to feel remorse for it.

2.7. Moral Wrongness and Feelings of Obligation

I will conclude this chapter by considering the relationship between the concept of moral wrongness and the con-attitude of feeling obligated not to do something. Applying the lessons we have learned about the relationship between ethical judgments and valenced attitudes to this particular case will be helpful for what is to come. The attitude or emotion of feeling obligated to do something has not been much discussed in recent literature. I think that we can characterize this emotion as a kind of prospective guilt-tinged aversion that one characteristically feels upon contemplating the prospect of performing or failing to perform an action, which performance or omission one takes to constitute doing something morally wrong. I think that Brandt (1959) gives a nice description of this emotion in his discussion of his "parked car episode":

> [I noticed] a car pulled off to the side of the road, with a man in it slumped over the wheel – as if asleep or ill…There was actually an impulse to stop…. In such cases, we normally say that we did not want to do a certain thing, but did it because we thought we ought. The writer was… not considering stopping in order to terminate some organic discomfort…Nor… in order to get or do something for himself…Some psychologists, at least partly in order to give recognition to the distinctiveness of the motivation, have suggested calling this sort of experience an "experience of requiredness"; but…there is no reason why we should not use an ordinary mode of speech to cover it: "I felt an obligation to…" (Brandt 1959, 116-118).

This feeling of obligation or prospective guilt-tinged aversion is also what Mill (1863, Chapter III, paragraphs 3 and 4) seemed to describe as an "internal sanction of duty…a feeling in our own mind… attendant on violation of duty, which in properly cultivated moral natures rises, in the more serious cases, into shrinking from it as an impossibility," and "a mass of feeling which must be broken through in order to do what violates our standard of right."

As associated as these feelings of obligation may be with judgments that one is morally obligated to do something, it is possible to feel obligated recalcitrantly, or to feel obligated not to do things that one judges not to be wrong. For example, a man from a background with restrictive views about sexual morality might feel obligated not to engage in certain sexual practices even though he now thinks them perfectly morally permissible. Or a woman in an abusive relationship might feel obligated not to leave her partner, but be thoroughly convinced that she is in no way morally obligated to stay with him. As with our discussion of recalcitrant guilt in Section 2.3, it seems that the man and woman could in this way recalcitrantly feel obligated without any of the conflicting inferential tendencies, views about appropriate conduct, or views about their own responses required for conflicting judgments about what they are morally obligated to do. All of these tendencies could be on the side of a judgment that they are not morally obligated to omit what they feel obligated to omit, and none on the side of a judgment that it would be morally wrong to omit it. Moreover, it seems that the man and woman could recalcitrantly feel obligated without any of the spontaneous appearances and tendencies to set burdens of proof in inquiry required for an intuition or sub-judgmental moral evaluation in conflict with their judgments about their moral obligations. Suspending judgment, their intuitive tendencies could be all on the side of seemings that they are not morally obligated to omit what they feel obligated to omit, and none on the side of intuitions that it would be morally wrong to omit these things.

Feelings of obligation or prospective guilt-tinged aversion thus seem explicable, not as states that essentially involve tokenings of the concepts MORAL OBLIGATION or WRONGNESS, but rather as manifestations of a phenomenal, motivational, and attention-directing syndrome. Among other things, feeling obligated to do something involves motivation to do it, and focused attention on how one can bring it about that one does it. It is very important to distinguish the motivations involved as components in feeling obligated to do things from motivations to do things in order to avoid feeling guilt, and indeed from motivations to stop oneself from having the feelings of obligation themselves. If one could take a pill that would prevent all unpleasant feelings of guilt for having done something, or indeed a pill that would terminate all uncomfortable feelings of obligation not to do it, the latter kinds of motivation not to do it in order to prevent or

terminate the feelings of guilt or obligation would be eliminated in favor of a motivation to take the pill. But the motivation not to do things involved as a component of feeling obligated not to do them would not be affected by the availability of such pills. When you feel obligated not to do something, you are motivated to avoid doing it either for its own sake (as when you feel obligated not to harm innocents), or as a way of avoiding doing something else that you feel obligated to avoid doing (as when you feel obligated not to fire rifles at innocents).

Thus, in addition to retrospective guilt we have an emotion of feeling obligated to do something, which involves motivation to perform the actions towards it is felt. But how exactly should this prospective guilt-tinged aversion to failing to do something, and its fittingness, figure into our understanding of concepts like MORAL WRONGNESS and BLAMEWORTHINESS? Above we saw the virtues of analyzing moral blameworthiness in terms of the fittingness of anger and *retrospective guilt*. But Gibbard (1990) crucially draws attention to some ways in which the concepts of moral blameworthiness and moral wrongness can come apart, including the following example:

> Imagine that in a paroxysm of grief I speak rudely to a friend who offers condolences, and so hurt his feelings. My rudeness is unprovoked, but understandable in the circumstances. I have thus acted wrongly, but because of my agitated state, it may not make sense to blame me (Gibbard 1990, 44).

Gibbard concludes that "we need a distinct concept of wrong…as opposed to blameworthy," noting that while the concept of blameworthiness is *retrospective* in character, the concept of wrongness is *prospective*. Unlike assessments of blameworthiness, assessments of moral wrongness are most intimately related to asking which of the acts open to oneself are wrong, and being motivated not to perform these. I think that the best way to understand this forward-looking character of the concept of moral wrongness is to see that it is concerned with the fittingness of prospective guilt-tinged aversion rather than retrospective guilt. Indeed, the same reasons that tell in favor of the above analysis of MORAL BLAMEWORTHINESS in terms of the fittingness of retrospective guilt and anger support the following analysis of MORAL WRONGNESS in terms of the fittingness of feelings of obligation or prospective guilt-tinged aversion:

Fitting Attitude Analysis of Moral Wrongness:

To judge that agent *A*'s act of ϕ-ing is morally wrong is to judge that it is fitting for *A* to feel obligated not to ϕ (or equivalently: to judge that it is fitting for *A* to feel prospective guilt-tinged aversion towards ϕ-ing).[79]

Like the above analysis of moral blameworthiness, this analysis has the virtue of explaining what is common to all coherent judgments about moral wrongness, which, like the corresponding judgments of blameworthiness we have observed, can attribute wrongness to wildly disparate acts. But this analysis can also help explain the gap Gibbard noted between simply judging an act wrongful and moreover judging it blameworthy. Combining this analysis of wrongness with our existing analysis of blameworthiness, we can understand, say, thinking that someone's lashing out in grief was wrong but not blameworthy as a thought to the effect that although it isn't fitting for us to feel angry at the person who lashed out and she shouldn't have to feel guilt for lashing out, it still was the case that before she lashed out she should have felt obligated not to do it.

A final virtue of this FA analysis of judgments about moral wrongness is that, given the influence that judging an attitude fitting has on our actually coming to have it, the analysis can explain how moral judgments can be motivating. It can explain how moral judgments motivate us, not only to try to make amends for blameworthy wrongs we take ourselves to have committed after the fact, but how they can moreover motivate us to refrain from committing what we take to be wrongs in the first place. Since prospective guilt-tinged aversion towards performing an act involves motivation not to perform it, judging that one should feel this guilt-tinged aversion towards performing an

[79] This is only a first-approximation to what I regard as an adequate fitting attitude analysis of moral wrongness. As we shall see in Chapter 4, the feelings of obligation are called for only if one is insufficiently motivated to omit the wrongful act, and the feelings are called for in a particular kind of way – as mandatory (like degrees of credence given one's evidence) rather than as simply justified (like feelings of anger in response to blameworthy actions).

act will tend to cause one to actually feel the aversion, including its motivation not to perform the act.[80]

[80] Gibbard (1990) gives an impressive expressivist semantics of judgments of warrant or fittingness reasons for attitudes that seeks to explain how such judgments have this kind of attitude guiding feature. One of the central initial motivations for expressivism was an attempt to explain how moral judgments can be motivating, and paired with his analysis of blameworthiness, Gibbard's expressivist semantics can admirably explain how judgments of blameworthiness can motivate making amends and punishing transgressors. Unfortunately Gibbard's expressivist semantics cannot combine with either his analysis of blameworthiness or the explicit analysis of wrongness in (Gibbard 1990) to similarly explain what is perhaps the most central case of moral motivation – that of an actor not to perform actions she deems to be wrong. But Gibbard can immediately remedy this defect by adopting the foregoing fitting-attitude analysis of wrongness, enabling the central case of motivation not to do what one judges to be wrong to be explained along the same lines as motivation to make amends for and punish what one judges to be blameworthy. (In fact, Gibbard (2006, 2007) seems to have taken steps along these lines, apparently taking a favorable attitude towards this analysis, or at least the use of prospective guilt-tinged aversion in the analysis of moral concepts).

Chapter 3

Fitting Attitudes and Reasons for Action

In the last chapter I argued that we should abandon the judgmentalist view that pro- and con-attitudes involve ethical evaluations. Rather, I contended, we should understand ethical evaluations themselves as evaluations of the fittingness of pro- and con-attitudes, where these attitudes are understood as syndromes of affect, motivation, and directed-attention, and we can understand assessments of these attitudes as fitting or warranted independently of the idea of ethical evaluation. Thus, we can understand the judgment that an outcome is good as a judgment that it is fitting for us to desire that outcome, we can understand the judgment that it is wrong to do something as a judgment that it is fitting to feel obligated not to do it, and so on.

Now, when judging that something is *F* involves judging that it is *G*, it is an analytic truth, or a truth guaranteed by the content of our concepts, that anything that is *F* is *G*. If believing someone to be a bachelor involves believing him to be male, then it's an analytic truth that anyone who is a bachelor is male. Similarly, if believing someone a vixen is nothing more than believing her to be a female fox, it's an analytic truth that someone is a vixen if and only if she's a female fox. Applying this to the present case, if ethical judgments can be analyzed as judgments about fitting attitudes, we will have an analytic equivalence between ethical facts and facts about the fittingness of valenced attitudes. Thus, if a state of affairs really is good, then it really is fitting to desire it; if an act really is morally wrong, then it really is fitting to feel obligated not to perform it, and so on. Because analyses of one kind of judgment into another in this way support analytic relationships between the facts that the judgments represent, I shall slide rather freely between talk about what it is to make a certain kind of judgment (e.g. "to judge an act blameworthy is to judge it to befit guilt and anger") and talk about analytic truths that

involve the facts represented by that judgment (e.g. "it's an analytic truth that if something is blameworthy, then it befits guilt and anger").[81]

Suppose, then, that a certain state of affairs – one, say, in which children in developing countries get enough to eat – is good. If the foregoing FA analyses are correct, it follows from this as an analytic truth that it is fitting for us to desire this state. Now, as we have seen, desiring a state of affairs involves being motivated to bring it about. But if we should desire that children in developing countries get enough to eat, and this desire that we should have involves being motivated to bring it about that these children get enough to eat, is it not the case that we should be motivated to bring this about? And if we should in fact be motivated to bring it about that the children get enough to eat, do we not have reason to bring this about? Or suppose that it would be morally wrong for me to do something – say to lie in a particular situation. According to the foregoing FA analyses, this analytically entails that I should feel obligated not to lie in that situation. As we have seen, feeling obligated not to do something involves being motivated not to do it. But if I should feel obligated not to lie in the situation, and this involves being motivated not to tell the lie, should I not also be motivated not to tell the lie? And if I really should be motivated not to lie, do I not also have reason not to lie?

I think that these kinds of facts about how to be motivated and what to do follow from ethical facts in the ways just suggested. But to make this out, we need to examine how in general thought about the fittingness of motives and motivational states relates to thought about what to do. This is what I propose to do in this chapter. The understanding at which we will arrive is that to think that you should do something is to think that your doing it will promote the ends at which you should aim, where the notion of an end at which you should aim is that of an end that it is fitting for you to be moved to bring about as an end in itself. My contention will be that this understanding of our judgments about what to do best explains how we are guided in decision and action by our basic, most fundamental deliberations about what is worth doing and why. This understanding of judgments about what to do will show how ethical thought, conceived

[81] This way of describing the analytic truths as involving the facts represented by normative judgments does presuppose descriptivism, but the analytic truths could easily be described in language neutral between the expressivist and descriptivist as truths involving the facts *signified* by the normative judgments, or as truths involving the *contents of* the normative judgments. I use the descriptivist talk about representation solely for the sake of simplicity.

of as thought about the fittingness of various states of intrinsic motivation, is at the heart of practical thought about what to do.

I will argue on this basis that the FA analyses of ethical concepts of the kind presented in chapter 2 capture the essence of the practical concerns embodied in our ethical thinking. We could, of course, use ethical language to express other concepts; we could use phrases like 'good outcome' and 'morally wrong' to express things like OUTCOME WE DESIRE and ACT WHICH IS CONDEMNED BY DTHAT[indicating some list of rules, like the ten commandments].[82] But then our ethical language would fail to express concepts that play the role in practical reasoning that is played by our serious ethical thinking. When we use our basic methods of normative inquiry to get at whether an outcome would be good or whether an act would be wrong in conjunction with a practical problem, we are trying to get at a particular kind of reason we might have to bring about the outcome or to refrain from performing the act. However confident you are that the outcomes supported by these reasons are the outcomes you desire, and however confident you are that the acts opposed by these reasons are the acts condemned by the list, it is a conceptually open question whether this is so. What one uses to fill the conceptual gap are one's reflective equilibrium methods of basic normative inquiry, and the FA analyses of ethical concepts best explain how this works.

There are several important distinctions within our thinking about what to do or what we have reason to do, which must be observed on pain of confusion. First, I should clarify that when I talk about someone's reasons for action, I am talking about her *normative* or *justifying* reasons for action; about considerations that count in favor of her performing some act. Reasons of this kind must be distinguished from an agent's *motivating* or *explanatory* reasons, or considerations that answer the question "why did she do that?" whether or not they contribute to justifying her doing it.

Second, there are two senses in which we can talk about someone's having normative reason to do something. The first is an "objective" sense in which what you have reason to do is what the facts of your case actually favor doing, whether or not you

[82] The notation 'dthat' followed by a clause in brackets indicating what is demonstrated to specify a demonstrative expression is due to Kaplan (1989).

have access to these facts. To adapt an example of Williams (1981), let us assume that you are in a situation in which the fact that a drink will kill you is a weighty reason not to drink it (you have a life well worth living), and that the fact that a drink is gin and tonic is a good reason to drink it (it has, after all, been a long day, and you would very much enjoy some gin and tonic). Now if the stuff in the glass before you is in fact petrol mixed with tonic water, then even if you have excellent evidence that it is gin mixed with tonic, you have no good reason in the objective sense to drink it and weighty reason in the objective sense not to drink it. The second sense in which we can talk about someone's having reason to do something is a "subjective" sense in which what you have reason to do is what your best evidence suggests is favored by the facts of your circumstance, whether or not your evidence is misleading. Thus, as you have excellent reason to believe that the stuff in the glass is gin mixed with tonic (as good, say, as ours usually is when we go to bars), you have good reason in the subjective sense to drink it, and no weighty reason in the subjective sense not to drink it.

The story about how we get from what we would have objective reason to do if the facts of our case were such and so to what we actually do have subjective reason to do in light of our evidence is one of the most important stories told by philosophers and economists. It is, as I see things, the story of decision theory. But this is not the story on which I will be focusing. The story that will be my concern here is that of what makes it the case that we have objective reason to do something if the facts do indeed turn out to be such and so, or what makes a consideration count as an objective reason to begin with. This story of objective reasons is at least as important as the story of decision theory for the task of figuring out what to do in our actual circumstances. For, suitably interpreted, decision theory is a theory of "relative rationality" only – it tells us how to get from objective reasons and evidence to subjective reasons. It takes our goals and beliefs as given, and does not presume to tell us what our goals or beliefs should be (other than that they obey certain formal constraints). We can interpret it either as a theory of what to do assuming we have the right goals and beliefs, or we can interpret it as a theory of instrumental rationality, where one can be instrumentally rational whether or not one has the goals and beliefs that one should have.[83] But however we interpret it, we still need a

[83] See Darwall (1983) and Gibbard (1998).

theory of what our goals and beliefs should be in order to draw upon decision theory to arrive at an account of what to do in our actual circumstances.

Of course, one theory of what goals to pursue and what beliefs to have is that pretty much anything goes – that any set of goals and beliefs that satisfies the formal constraints of decision theory is just as rational or just as well supported by reasons as any other. A similar view is that all talk of substantive rationality is nonsense and that the only questions about what to do that make any sense are questions of instrumental rationality. But these are radical and controversial views – far more radical and controversial than anything that shows up in decision theory proper. One must be careful not to besmirch the good name of decision theory by treating it as a full theory of how to act that is inextricably wedded to these conjectures about substantive rationality. Similarly, one must be careful not to suggest that these conjectures are supported by the same evidence or carry the same kind of authority as the good and careful work done by decision theorists.

3.1. Analytic Humeanism

Controversial conjecture though it may be, something like the "anything goes" theory about what goals to pursue holds sway as a theory that has some plausibility under the banner of the so-called "Humean theory of practical reasons." According to this view, what one has objective reason to do is whatever will satisfy one's actual intrinsic desires – whatever will, that is, bring about what one wants for its own sake. Now there might be such a concept as that of an INSTRUMENTAL REASON TO DO SOMETHING, where having instrumental reason to do something is just a matter of that thing's satisfying one's given ends. We might think, however, that the most important concept of an objective reason to do something is distinct from that of an instrumental reason to do it. The versions of the Humean theory of practical reasons that will be interesting for our purposes are those that go beyond offering their view as a theory of what we have instrumental reason to do. These versions will allow a distinction in principle between what we have instrumental reason to do and what we actually have objective reason to do. One interesting version

would go on to maintain that what we have objective (and not merely instrumental) reason to do is whatever will satisfy our actual desires. Another would offer Humeanism as a theory of what we have instrumental reason to do, go on to claim that thought and talk about our actual objective reasons as distinct from our instrumental reasons makes no sense, and conclude that instrumental reasons (of which Humeanism is the true theory) are the only important normative reasons for action there are.[84]

Whether any interesting version of the Humean theory of practical reasons could be *true* is an interesting question on which our work here shall eventually bear. But what I should like to consider now is whether the Humean might try to account for *what we are doing* when we think about what we have reason to do. The question I wish to explore in the first instance is not the substantive normative question of what we actually have reason to do, or which considerations count in favor of performing which actions (and how strongly they do so). It is rather the metanormative question: what is it to judge that you or someone else has reason to do something? What is it to judge that a consideration counts in favor of doing something?

A natural answer to these metanormative questions might well be drawn from the Humean theory of practical reasons. Perhaps judgments about what someone should do are nothing other than judgments about what would satisfy her intrinsic desires. Thus, the suggestion would go, to think that Smith should take an umbrella is just to think that Smith's taking an umbrella will satisfy his intrinsic desires (which might include a desire to stay warm and dry). Similarly, the Humean suggestion would go, the judgment that consideration *R* is a normative reason for Smith to do *X* is nothing other than a judgment that *R* contributes to making it the case that Smith's doing *X* will satisfy his actual intrinsic desires. To think that the fact that it is raining is a reason for Smith to take an umbrella is to think that the fact that it is raining helps make it the case that Smith's taking an umbrella will satisfy his actual intrinsic desires. This version of the Humean theory of practical reasons thus offers itself as an analysis of our judgments about what to do; according to this theory it is actually an analytic truth that one has reason to do

[84] The distinction between these two interesting versions will not be very important for our purposes. For convenience I will mostly talk in terms of the first – the kind that takes talk about what we actually have objective reason to do on board and offers Humeanism as the true theory of that. But what I say could easily be put in terms of the other version that is eliminativist about actual objective normative reasons and contends that instrumental reasons are the only reasons worth thinking about.

whatever will satisfy one's actual intrinsic desires. I will therefore refer to this theory as 'Analytic Humeanism'.

I actually think that there is quite a bit to be said in favor of Analytic Humeanism, some of which I will discuss in more detail in the final chapter of this dissertation. One cardinal virtue of the theory is that it looks like a plausible way to explain how assessments of what an agent has normative reason to do differ from other kinds of normative assessments. To judge that an entity has reason to do something is different from simply judging that its doing it would be good or something we should hope for or promote. When volcanoes fail to erupt and kill people, they do something that we should hope for and (to the extent we can) promote, but it would be absurd to think that they do something that they had reason to do. Similarly, when in fair competition with another agent that other agent does something that gives her an advantage, she does something that one may have reason to hope she doesn't do and try to prevent her from doing, but nonetheless does something that she has reason to do. The Analytic Humean offers a compelling explanation of this. It is incoherent to think that volcanoes are as we take them to be (with no mental lives at all) and to say that they have normative reasons to do things because the very idea of having normative reason to do something is that of the action's helping to satisfy one's desires. Similarly, in cases of competition we have reason to do what the other agent has reason to prevent because when we both want to win, the satisfaction of one person's desire must frustrate the satisfaction of the other's.

Another attractive feature of Analytic Humeanism is its apparent ability to explain how thought about one's normative reasons can be both descriptive and essentially action guiding. As we saw in Chapter 1, one of the central, puzzling features about normative judgments is that there seem to be normative facts of the matter that we can hook onto via our methods of normative inquiry, but that our normative judgments seem to play a more intimate role in guiding our responses than any other kind of descriptive judgments, or judgments that purport to represent facts. Now, whether an act will satisfy one's intrinsic desires is a matter of descriptive fact that one can hope to hook onto through inquiry into such matters. But at the same time, it is simply part of what it is to desire that *P* for one's judgments that doing *X* will bring it about that *P* to give rise to motivation to do *X*.[85]

[85] See Stalnaker (1984)

Now as we saw in chapter 2, there seems to be a sense of 'desire' in which desires are more than motivational states; to desire an outcome in this sense requires that one have the right kind of positive affect towards it, that one exhibit the right patterns of directed attention, and so on. There is, however, also a thinner sense of desire, in which to desire an outcome is simply to be motivated to bring it about, or to be in the kind of state that combines with representations that doing *X* will bring it about to give rise to one's doing *X*. It seems quite clear that versions of Analytic Humeanism that are formulated in terms of the thinner notion of intrinsic desires as merely states of intrinsic motivation are far more plausible. For why should we discriminate against intrinsic motives to do things just because we fail to have positive affect towards their objects? In our above example of Smith who "wants to stay dry," his motivation may be more of an aversion to getting wet: it is more like he has negative affect to states of wetness than any positive affect towards states of dryness. But surely it would be strange to say that Smith has reason to take an umbrella if he looks on keeping dry in a favorable way, yet no reason at all if he simply looks on getting wet in a disfavorable way (and stranger still to say that this is nothing less than an analytic truth!). Similarly, if the Humean wants to say that an agent has reason to bring about the outcome she most favors when all of her options will bring about something she favors, the Humean had better be prepared to say that the agent has reason to bring about the option she least disfavors when all of her options will bring about something that she disfavors.

Another important thing to clarify about Analytic Humeanism concerns the kinds of things it allows to be the objects of the intrinsic motives that we are supposed to have reason to satisfy. For convenience I have so far formulated Analytic Humeanism as saying that we have reason to do what satisfies our intrinsic motives to bring about certain states of affairs (like our being dry). But as we have seen not all intrinsic motivations are motivations to bring about states of affairs. Most of us, for instance, feel intrinsically obligated not to kill: we have an intrinsic guilt-tinged aversion toward killing (when we take death to be bad for the victim). This is not simply an aversion to there being killing in the world – faced with the option to kill one person in order to prevent two killings, our motivation not to kill would tend to make us refrain from killing the one.

Nor is ours an aversion to its being the case that we kill ourselves. Consider, for instance, the following variation of the classic organ transplant case (in which a surgeon can kill one healthy person to transplant his organs into five others) due to Thomson (1985):

> Suppose that what had happened was this: The surgeon was financially badly overextended last fall, he had known he was named a beneficiary in his five patients' wills, and it swept over him one day to give them chemical *X* to kill them ["Now chemical *X* works differently in different people. In some it causes lung failure, in others kidney failure, in others heart failure. So these five patients who now need parts" need them because the surgeon intentionally gave them the chemical to kill them for their inheritances]. Now he repents, and would save them if he could. If he does not save them, he will positively have murdered them. Does *that* fact make it permissible for him to cut the young man up and distribute his parts to the five who need them? (Thomson 1985, 98).

What matters for our purposes is not whether it is permissible to cut up the young man; what matters is that our aversion to killing would motivate us not to do so if we were in the surgeon's position, whereas a mere motivation to bring it about that we not kill would not incline us against killing the one over preventing ourselves from having killed the five.

Finally, our aversion to killing is not even an aversion its being the case that we kill at the time of decision. The latter would motivate us to do things to prevent ourselves form killing at the time of decision other than simply not killing when the time of decision comes. It would, for instance, motivate us to try to make ourselves into the kind of people who will not kill at the moment of decision, or perhaps motive us to stay away from situations in which we will be likely to kill. But even if some of us have these kinds of motives, it is important to observe that they are distinct form our guilt-tinged aversion to killing itself, which one can have in their absence. Someone convinced of the truth of direct consequentialism might well have no motivation to make himself into the kind of person who refuses to kill one to save five, even though he retains a strong (though recalcitrant) aversion to killing that would motivate him not to kill the one. Similarly, a staunch deontologist might retain his aversion to killing but have no reluctance to entering a situation in which he knows that he will be tempted to kill, and have no inclination to send someone else in his stead, simply on the grounds that his killing at the time of decision is something that might happen. Our aversion to killing really is distinct

from an aversion to dirtying our hands by being the one who does the killing (though some people no doubt have that too).

Our aversion to killing then, cannot be reduced to a motivation to bring about a state of affairs. I think that similar remarks could be made for many other kinds of motives, including those we usually have not to do what we regard as shameful or lowly, to make amends when we feel guilty, to behave punitively when we feel anger, and those we often have to take walks, to dance, and to sing. Motivations like these are all best treated as having actions, rather than states of affairs, as their objects – they are motivations *to do* or *refrain from doing things* rather than motivations *to bring about* a kind of outcome.[86] The question for the Analytic Humean, then, is whether she thinks that the judgment A HAS REASON TO DO *X* is a judgment to the effect that doing *X* will satisfy all of *A*'s intrinsic motives, including *A*'s intrinsic motives to do or refrain from doing things, or whether she thinks it is a judgment that doing *X* will bring about those states of affairs that *A* is motivated to bring about.

If the spirit of the Humean theory is to take our ends or goals as given, or as they actually are, then clearly the Analytic Humean should make the former kind of identification. For people who are intrinsically motivated to do things have doing those things as final ends in the same way that people who are intrinsically motivated to bring about states of affairs have as final ends the bringing about of those states. When you are intrinsically motivated to bring about an outcome, you aim at that outcome on its own account and independently of its further effects. Just so, when you are intrinsically motivated to do (or refrain from doing) something, you aim at doing (or omitting) it on its own account and independent of its further effects. Of course, you might be intrinsically motivated to (do or omit) something without in some sense "adopting" your doing it as "your aim" – without, that is, committing to doing it or forming an intention to do it. But the exact same is true of desiring or being motivated to bring about a state of affairs – this too is not yet to intend, commit to, or adopt as your aim the bringing about of the state.

It would, indeed, appear to be rather arbitrary for the Humean to restrict our reasons to do what we aim at doing to our aims to bring about states of affairs. Why is it that being intrinsically motivated to bring about a state of affairs makes it the case that

[86] I am grateful to Jonathan Dancy and Douglas Portmore for very helpful discussion on these points.

one has reason to bring it about, but being intrinsically motivated to do something gives rise to no corresponding reason to do it? A Humean might, I suppose, think that there is just something more rational about the sort of direct consequentialist reasoning that looks always to bringing about certain states of the world rather than to doing or omitting certain things on their own account. But I cannot see why this would be at plausible if we are taking our actual ends as given – why people who act on their desires to bring about as much suffering as possible are somehow more rational than people who act on their desires to take walks. It would, in any event, seem to be a grave mistake for the Analytic Humean to rule out non-(direct)-consequentialist thinking about what one should do as incoherent. Of course, the only alternative is for the Analytic Humean to allow that anyone who has intrinsic motives to do things actually has reasons to do them that are not just reasons to bring about certain states of affairs. But that is the price one pays for offering such a radically subjectivist theory as an analysis of our thinking about what to do.

3.2. The Shortcomings of Analytic Humeanism

The version of the Analytic Humean theory at which we have arrived is thus that to judge that *A* has reason to do *X* is to judge that *A*'s doing *X* will contribute to the realization of the objects of her intrinsic motives – to bring about some state that *A* is intrinsically motivated to bring about, or to help make it the case that *A* does (or omits) something that *A* is intrinsically motivated to do (or omit). More generally, this theory maintains that to judge that *A* has most reason to do *X* is to judge that *A*'s doing *X* will realize the objects of A's strongest intrinsic motives (where an intrinsic motive to do *X* is stronger than an intrinsic motive to do *Y* if it causes you to try to do *X* instead of *Y* when you take it that you can do either but not both).

Now it seems plain that this theory flies in the face of much of how we usually seem to think about reasons for action. For it seems that people can be intrinsically motivated to do all kinds of crazy things. A person might, for instance, be intrinsically moved to collect blades of grass, count the number of bottle-caps in the world, turn on

radios, wash his hands frequently, avoid harmless spiders, beat-up people who look like his father, or what have you. Indeed, the psychological principles of classical and operant conditioning would seem to suggest that by associating the right stimuli or punishing and rewarding the subject's responses we could induce in him almost any bizarre intrinsic motive we like.[87] It certainly seems coherent (as well as true) to say that the people who have these motives have no reason to do what would realize their objects.[88] Of course, a person with these motives might have reason to do what satisfies them as a means to the end of avoiding the pain or frustration that might accompany not satisfying them. But what the Humean theory affirms, and common sense seems to deny, is that people with these motives have reason to do what will realize their objects simply on account of the fact that they have these motives and that performing the relevant acts will realize their objects. According to the Humean theory, they have reason to perform these acts whether or not (or at any rate in no way because of the fact that) their performing them will spare them pain and frustration.

Similarly, it seems that people can lack motivations to do all sorts of things that we can coherently (and plausibly) think that they should do nonetheless. A depressed person might have no motivation to live, a short-sighted person might have no motivation to take care of his health or save for his future, and an unfeeling person might have no motivation to refrain from torturing innocent beings for fun. But it certainly seems coherent (as well as true) to say that the people who lack these motives still have reason to do these things, and that these reasons can be conclusive whether or not doing these things will serve the agent's other intrinsic motives. Thus, if a person's strongest intrinsic motive is to spend her days counting blades of grass or to torture innocents for fun, common sense maintains that it is not only coherent but really quite plausible to say that the person has most reason not to do what would realize the object of her strongest intrinsic motive.

---

[87] On this topic see Mill (1863) and Darwall (2002).

[88] Or no reason to do what would realize the objects of these motives simply on that account. It may be that actions that would realize the objects of motives that are intuitively crazy would also help make it the case that an agent does what she intuitively has reason to do. Thus, someone might have a pathological motivation to wash his hands a lot but care nothing about his health. If there are lots of diseases going around, the person would intuitively have reason to wash his hands frequently, though it is on account of his health (about which he cares nothing) instead of his motivation to wash his hands.

But the proponent of Analytic Humeanism has an important response to these intuitions about coherence (and plausibility). She can, I think with some credibility, try to debunk these intuitions as stemming from a confusion between judgments about what someone should do on the one hand and other kinds of normative assessments of her actions on the other. The Analytic Humean can return us to the distinction we mentioned earlier between:

(a) an entity's doing something the occurrence of which we should hope it doesn't do and oppose if we can, and

(b) an entity's doing something that it has most normative reason not to do.

As we observed, the occurrence of natural disasters are instances of (a) but not (b), while the sub-optimal play of our opponents in fair competitions is an instance of (b) but not (a). The Humean can thus suggest that the above intuitions that agents should not do what satisfies their strongest intrinsic desires to spend their days counting blades of grass, to torture innocents, or whatever, are actually confusions of (a) with (b). Sure enough, we have reason to try to get people not to spend all their time counting grass and we have reason not to let them torture innocents – you and I care about the people in question and we know that everyone will be better off if they don't do these things. But when we go on to say that the people in question actually have most reason not to fulfill these desires, we are confusing, perhaps out of wishful thinking, what we have reason to prevent these people from doing with what they have reason not to do.

In support of the view that we have made this confusion, the Humean might put to us the following question. Why, in any given case, should we think that someone's doing something like torturing innocents for fun is not only something that we should hope she doesn't do and prevent her from doing, but moreover something that she has reason not to do? Is it not that, above and beyond our having reason to oppose her action, we think that she could correctly reason her way to refraining from performing it? Correct practical reasoning, however, is not just any process by which an entity comes to do or be as we might want her to do or be. Were we to neurally alter a shark so that he no longer attacks us, he would not thereby have correctly reasoned himself to refraining from doing so. What, then, is correct practical reasoning? A natural suggestion, which we might trace

all the way back to Aristotle (ca. 350 B.C.E.), it is that correct practical reasoning is correct instrumental reasoning, or a matter of correctly determining which actions will best promote our given ends. If this is all there is to practical reasoning, and our actual ends are those embodied in the objects of our intrinsic motives, then the only thing anyone could correctly reason her way to doing is what will realize the objects of her strongest intrinsic motives.

Underlying the Analytic Humean's argument is an idea about the relationship between practical reasons and correct practical reasoning that seems to capture nicely the distinction we want between thinking merely that we have reason to oppose someone's action and thinking moreover that her action is something that she has reason not to perform.[89] The challenge to the opponent of Analytic Humeanism is thus to show how practical reasoning or inquiry into what to do can involve more than simply determining what will realize the objects of one's current intrinsic motives. One claim along these lines is that practical reasoning can involve a kind of vivid appreciation of considerations and an imaginative simulation of circumstances that is not simply a matter of determining what will satisfy one's existing intrinsic motives. Darwall's (1983, 30-41) example of Roberta has also been taken as an example of coming to appreciate practical reasons in this kind of way. Roberta, as Darwall tells the story, grows up in a relatively sheltered environment, and upon going to university she sees a film that vividly portrays the plight of textile workers in the southern United States. After seeing the film she decides to spend some time each week helping in a boycott to improve the conditions of the textile workers. Darwall claims that this decision is the result of genuinely new intrinsic motives to alleviate the workers' suffering as a result of her appreciation of their plight.

[89] I should emphasize, moreover, that this way of marking off our concept of what someone has reason to do in no requires thinking that one must actually reason about what to do in order to do what one has reason to do. All that is required is that a process of correct reasoning could in principle support the action in question; in practice this reasoning might be unavailable and our actual reasoning might inevitably lead us astray. The view expressed by Tennyson's words about members of the Light Brigade that 'Theirs [was] not to reason why / theirs [was] but to do and die' is perfectly consistent with this way of drawing the distinction (whether read as a view about subjective or objective reasons). So too is the view that the correct process of reasoning that would reveal that the members of the light brigade had objective reason to obey orders is quite simple, perhaps consisting of a little more than a direct revelation of intuition (plus perhaps inference on the basis of this intuition and at most the ruling out of attempts to debunk the intuition).

There are, however, several questions that Darwall's example raises. The first is why exactly we should treat Roberta's development of new desires as an instance of practical reasoning or a response to (what she takes to be) practical reasons.[90] Coming to have new motives as a result of seeing a film might look a lot like a "conversion experience," or an episode that alters one's motivations and consequent actions in a way that cannot be treated as a response to what one took to be justifying considerations before the fact. The use of vivid images in 20th century propaganda might make us particularly skeptical of this feature of the example. *Birth of a Nation*, *Triumph of the Will*, and *The Eternal Jew* probably instilled new desires that motivated a great deal of decision and action, but we might hesitate to view the inculcation of the new desires as a response to practical reasoning, however shoddy. Vivid images in film might alter our motives in an un-reasoning way – much like classical and operant conditioning, only quicker.[91]

The second question that Darwall's example raises is, to the extent that Roberta's vivid imagining *was* an exercise of practical reasoning, how far it actually went beyond an attempt to figure out what would realize the objects of her existing motives. To get around the previous problem about possible un-reasoning alterations of desires, consider a case of consulting vivid representations that is more obviously an example of practical reasoning. In trying to decide whether to take, say, a certain job, one naturally seeks to vividly imagine and represent to oneself what it will be like to live under those conditions of employment. It is, however, somewhat unclear to what extent this is more than a fancy form of instrumental reasoning that goes beyond anything that could be expressed in something as crude as a spoken language. We may already be motivated to have or avoid having experiences with a certain phenomenal character, and to live in social arrangements that have certain properties, but these features of experience and social arrangements may be largely ineffable. As such, our best access to whether these features will be present in a situation – and thus whether the situation will realize the objects of our motives – may well be to vividly simulate the situation.

[90] Cf. Finlay (2007).

[91] Though it might be rather like the operation of so called "bait shyness," in which we become aversed to a kind of food after only one subsequent bout of nausea (see e.g. Zimbardo and Weber 1997).

In the same way, Roberta might have standing motives to help people when she can. Darwall does well to emphasize that Roberta might not be motivated by either an occurrent desire or an abstract motive to prevent suffering. But *our* Analytic Humean is quite ecumenical about which intrinsic motivational states are apt for entry into instrumental reasoning: desires, motivations to perform certain acts, occurrent, dispositional – it's all the same to him. What is not obvious is why Roberta might not have rather strong motives with relatively ineffable content – say a dispositional motive not to let anyone she runs across experience dthat[indicating some phenomenal character, social relationship, or simulation thereof]. If she does, then her best access as to what would satisfy such a motive might well be the vivid information presented by the film. The Humean might try to claim that, to the extent that the vivid images are playing a role in Roberta's practical reasoning – the kind of process undergone by our prospective employee as opposed the kind of process undergone by a teenager viewing a propaganda film in the first half of the 20th century – they are simply conveying information about what will satisfy existing motives with this kind of complex and dispositional content.

The third question raised by Darwall's example of Roberta is, to the extent that coming to have new desires as a result of vividly appreciating something *does* seem to be an exercise of practical reasoning, what exactly makes it so. Brandt (1979) presented compelling examples of intuitively irrational motives that could be de-conditioned by sufficiently vivid appreciation of - or repeated exposure to - accurate information. He gives the example of Albert, who has an intrinsic aversion to being in the presence of rabbits, simply because someone produced a loud noise in his vicinity when he was reaching out to interact with a rabbit (Brandt 1979 11-12). Brandt notes that repeated and vivid attention to the fact that there is no causal connection between rabbits and loud noises could "disabuse" Albert of the aversion.

Motivated by examples such as this, Brandt valiantly attempted to give a reforming account of our practical concerns as the realization of the objects of those intrinsic motives that could survive maximal and vivid confrontation with the facts, which Brandt termed "cognitive psychotherapy." The problem, however, was that in general the idea of what would serve those of our ends that are maximally resistant to de-conditioning seemed to fail rather badly to align with what we are really after in

deliberation about what to do. The mere fact that certain of our desires are so ingrained and recalcitrant that they would fail to extinguish under cognitive psychotherapy (like some phobic desires might be) seems to be insufficient reason for us to decide to "go with them" and seek to realize their objects. Similarly, the knowledge that certain of our desires, like certain desires not to take bribes or to refrain from killing ourselves,[92] might extinguish under maximal and repeated exposure to certain pieces of information seems to be insufficient reason for us to decide to abjure their objects as in no way worth realizing.

In the final analysis, what seems to account for both the plausibility of Brandt's motivating cases and the inadequacy of resistance to de-conditioning as the sole criterion of rationality is this. What really gives us confidence that motives like Albert's are irrational are our substantive intuitions about their irrationality rather than our suspicion that they could easily be extinguished by vivid exposure to information. If we think that a certain fact counts against our being motivated to do something, then we will treat the extinction of that motive in light of our vivid appreciation of that fact as a response to practical reasoning. But we will think this only because of our prior commitment to the idea that the fact is a consideration in light of which the motive is inappropriate. For suppose we are convinced that facts about what we could do with the bribe-money do not show our aversion to taking the bribe to be inappropriate. We then will not take the extinction of our aversion to taking the bribe in response to the vivid presentation of facts about what we could do with the money to be an instance of our motives responding to our practical reasoning. This seems to suggest that if there is a kind of practical reasoning that guides the revision of our motives, vivid attention to considerations does little on its own to explain how it works, since any role it has seems to take facts about which considerations count in favor of which motives for granted.

But Brandt's account did seem to reflect a crucial aspect of what we are after in our practical reasoning. Intuitively, deliberation about what to do does not seem to be exhausted by attempts to determine what will satisfy our existing intrinsic motives, whatever they may be. We seem to be able to evaluate our intrinsic motivations

[92] For examples of this kind see Gibbard (1990, 20-21).

themselves as rational or irrational, and we take ourselves to have reason to do what will satisfy only those motivations that we deem rational.[93] Moreover, we seem to have a method of determining which non-instrumental motives are rational, and our judgments about their rationality tends to influence which non-instrumental motives we actually have.

Consider, for example, two different ways in which an agent might become convinced that she should eat a vegan diet. One way is for her to start off with a strong aversion to contributing to the suffering and death of animals, and to become informed that animals in the dairy and egg industries are routinely abused and killed very early in their lives after they exceed peak productivity.[94] But a second way for is for her to start off knowing about the conditions of the animals, and to become convinced by philosophical arguments that she *should* be averse to contributing to the suffering and death of non-human animals. These might appeal, for instance, to intuitions about how one should treat space aliens with the same psychology as humans, how one should treat humans with mental lives comparable to non-human animals, and whether a being's species membership independent of her psychology should make a difference to how one treats her.[95]

While the first way of being convinced to eat a vegan diet might involve nothing more than discovering what would satisfy one's existing intrinsic motives, the second way seems to go beyond this. It seems to involve a kind of inquiry into what one *should* be intrinsically motivated to do, which works by seeking out a reflective equilibrium among one's various intuitions about what kinds of things to avoid doing, what affects one's reasons to do things, general principles about what to do, and what one should do in particular cases. When as a result of this process we judge that we should pursue a new end like avoiding harm to non-human animals, our judgment tends to give rise to new intrinsic motives like aversion to harming animals for their own sakes.

I will consider three attempts to reconcile Analytic Humeanism with the kind of reflective equilibrium inquiry at work in the second way of being convinced to be vegan.

---

[93] Much as we observed above, even when we think that we should satisfy an irrational motive in order to avoid the irksomeness or unpleasantness of its remaining unsatisfied, we take our act to serve a perfectly rational motive to avoid annoyance or unpleasantness.

[94] See for instance Mason and Singer (1990, especially 5-6, 39-40 and 10-14).

[95] See for instance McMahan (2002, 2003)

While I will argue that all three of them fail, I think that their failures bring out important features of practical reasoning that an adequate theory must explain.

First, the Humean might try to explain the kind of attitude revision at work in the second way of becoming convinced and caused to be vegan in terms of something like a higher-order desire to conform one's motives to reflective equilibrium methods. Surely, the Humean might say, we desire to have those motives that reflective equilibrium methods prescribe. In the second way of being convinced and caused to be vegan, it is this desire, plus a belief that reflective equilibrium methods prescribe aversion to harming animals that explains the generation of the aversion.

The basic problem with this response is that our agent's aversion to harming animals can be generated more directly than the explanation in terms of a higher-order desire would allow. In general, a motivation to bring about an outcome can only bring it about by getting one to *do* things that bring it about.[96] This seems to follow simply from the functional role of motivations as states that combine with beliefs to directly produce *action*, as depicted below in Figure 4.

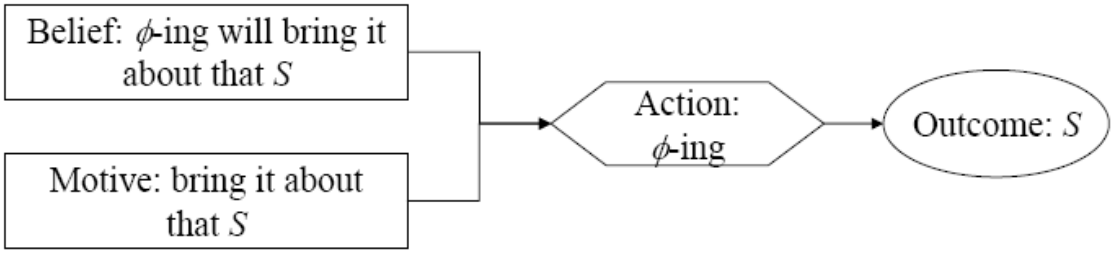

Figure 4: How beliefs and motives combine to produce outcomes

Hence, a motivation to have a new motive can only cause one to have it by causing one to do things to get oneself to have it. Such actions might include taking pills, classically conditioning oneself, or paying selective attention to certain things.

Thus, a desire to have the motives reflective equilibrium methods prescribe and a belief that they prescribe aversion to harming animals could only cause the aversion by

[96] That is, absent auxiliary apparatus like other people reading our minds and bringing about what they see we're motivated to bring about.

causing one to do these kinds of things to get oneself to have it. But this is not how philosophical reasoning typically guides our motives. Coming to the conclusion that one should, for instance, avoid harming animals as an end in itself can *directly* cause aversion to harming them without the mediation of actions undertaken to get oneself to have this aversion. This kind of direct influence is that of the judgment that a motivation is fitting on our coming to have the motivation, which we have seen before. This influence, recall, is symmetric to the direct influence of judgments about evidence on beliefs, and judgments about what to believe can be contrasted with desires to have beliefs in the same way judgments that motivations are fitting can be contrasted with desires to have motives. Judging that one's evidence supports believing that there are no deities can directly cause one to believe that there are none without one's having to do anything to get oneself to believe this. But a mere desire to believe that there are deities cannot cause theistic belief without first causing one to do things to bring the belief about.

A second attempt by the Humean to account for our agent's being convinced to be vegan by reflective equilibrium methods might contend that the agent discovered a strong aversion to harming animals that she had all along. Before the reflective equilibrium exercise, the agent simply did not realize that she had this aversion. What her use of reflective equilibrium methods did was teach her about the existence of this aversion already present in herself.

The basic problem with this response is that coming to judge that one should be averse to harming animals is causally efficacious in a way that coming to judge that one is averse to harming them is not. In general, one's motivations cause one to do what they are *actually* motivations to do, *not* what one simply thinks they are motivations to do. One might, for instance, have a desire to approach someone sitting at a bar to whom one is attracted, which one mistakes for a desire for beer. In such a case, one will expect that if getting beer and approaching the person come apart, one will do what procures beer rather than what gets one close to the person. But if one is actually motivated to approach the person rather than procure beer, the motive will (*ceteris paribus*) cause one to violate one's expectations and do what brings one near the person rather than what

procures beer.[97] If the person leaves the bar for the sandwich shop, one may find oneself doing so as well, even though one expected that one would stay and drink.

Thus, if prior to philosophical inquiry an agent had a strong aversion to contributing to harming animals of which she was unaware, this motive would *already* have combined with her belief that non-veganism contributes to their harm and caused her to eat a vegan diet. If all reflective equilibrium inquiry did was make the agent aware of her motive, it would not change her from non-vegan to vegan; it would simply change her from a vegan with a poorer understanding of her behavior to a vegan with a better understanding of it. But reflective equilibrium inquiry can make vegans out of non-vegans. In general, when philosophical inquiry causes one to think that one should do something, it tends to supply new motivation to do it.

A final Humean attempt to account for what goes on in the second way of being convinced and caused to be vegan might be to claim that the agent is simply moved by something like a first-order desire to *do* whatever reflective equilibrium methods prescribe she do. What she learns by reflective equilibrium methods is just that these methods prescribe that she *not do* what contributes to harming animals, and this belief combines with her desire to do what reflective equilibrium methods prescribe to cause her to eat a vegan diet.

---

[97] I should stress that this causal propensity will determine what one does when all else is held equal. One can of course have countervailing motives that prevent one's procuring beer. One can also have motives like those to avoid the frustration of unfulfilled desires, which will cause different actions depending upon one's views about what one desires. One can similarly take evidence about one's desires as evidence about what one would enjoy, which, in combination with motives to do what one would enjoy, can cause action in a way that depends upon one's views about what one desires. These are not, however, instances of action produced by the mere combination of beliefs that one is motivated to bring something about and beliefs that a certain action will bring it about. Rather, in these cases motives to avoid frustration or procure enjoyment *themselves* (and *not* beliefs about them) combine with beliefs about what will bring about *their* satisfaction to produce action.

It also seems plausible that there is a mechanism that causes some of our motives to conform to our theories or narratives about the kinds of motives we have. In some cases this might generate (or strengthen) motives like desires for beer and eliminate (or weaken) motives like desires to approach a person to whom one is attracted. I think that this process is the result of our accepting norms that cause our motives to conform to our views about what would make them intelligible, and I suspect that this is much of what is correct about accounts of practical reason like that of Velleman (2000) and perhaps also Korsgaard (1996). As we will see in Chapter 6, the acceptance of a norm for motivation is a distinct kind of mental state that can, unlike a desire to have a motive, directly influence the motives we have.

The first problem with this account is that, rather than motivating the agent to do one thing or another simply as means to the end of conforming to the dictates of reflective equilibrium methods, the agent's philosophical inquiry seems to alter her intrinsic motives themselves. Philosophical arguments for veganism of the kind we mentioned contend, for instance, that just as we should be non-instrumentally averse to harming mentally disabled humans, so too we should be averse to harming mentally comparable animals for their own sakes. Being convinced by this kind of argument tends to directly generate non-instrumental aversion to harming the animals.

To deny this intuitive picture seems to paint philosophical inquiry as necessarily producing the wrong kind of motives, or those Williams (1976) might object to as involving "one thought too many." It seems, for instance, that a philosophical argument against racism can convince someone to improve her treatment of members of other races. Such an argument might examine what race membership really comes to (e.g. skin color, facial features, area of ancestry), elicit intuitions that these things severally look normatively irrelevant (which might draw on intuitions about cases that isolate them), and seek to show that any intuitions that race is relevant are due to the conflation of membership in the same race with somewhat correlated features like degree of personal contact.[98] But something would surely be amiss if this kind of argument convinced someone to treat members of other races better simply as a means of doing whatever philosophical methods prescribe. Surely the case against racism purports to show that the person should improve her treatment of members of other races for their own sakes, not as just as raw materials for the enactment of the dictates of reflective equilibrium methods!

[98] Or, perhaps the racist has conflated the apparent relevance of race with the apparent relevance of some less savory features that the racist might only take to be correlated with race membership, like low intelligence and poor work ethic. Whether these features are so correlated with race would be irrelevant to the point at issue of race membership *per se* matters to how you ought to treat people. Even so, anyone who seems to think that intelligence and work ethic could matter in the way the racist seems to think race matters to how one ought treat people should presumably consider the philosophical arguments against this, such as those against such relevance of intelligence presented by Singer (1975). Of course, a racist will not be likely to treat things like intelligence and work ethic as mattering so much to how she ought to treat members of her own race, and the best explanation of her thinking they matter to how she ought to treat members of other races will presumably involve their role as mere attempts to rationalize a more obviously unjustifiable attitude towards them.

An identical problem befalls a similar Humean attempt to account for the influence of our reflective-equilibrium methods in terms of changes in moral belief combined with a *de dicto* desire to be moral, or to avoid doing whatever is morally wrong. As Smith (1994) observes, concern with doing what is moral *per se* seems more like a fetish than an appropriate response to our moral reasons. Thus, we seem to be moved to avoid harming humans with mental lives comparable to non-human animals for their own sakes rather than simply as a way to avoid doing something that would count as wrong; we seem to be moved to avoid harming them for their own sakes. The philosophical argument for veganism purports to show us that we have the same kind of reasons to avoid harming non-human animals, and being convinced by it seems capable of causing aversion to harming them for their own sakes rather than simply aversion to harming them as a way of doing something that is wrong. In the same way the argument against racism seeks to show that one has reason to treat members of other races better for their sakes rather than for the sake of avoiding wrongness *per se,* and being convinced by it can cause such intrinsic aversion.

Another problem for the Humean attempt to account for the influence of the above kind of philosophical reflection on our aversions and behavior in terms of a *de dicto* desire to avoid doing what is wrong is that the same kinds of reflection seem to alter our intrinsic motives in many non-moral domains. For instance, Nozick's (1974) experience machine argument can strengthen our motivations to pursue things like genuine achievement and knowledge for their own sakes, not just as means of doing what reflective equilibrium methods tell us to do. A similar Humean gambit might be to posit a *de dicto* desire to pursue what is good, but similar "one thought too many" objections would apply to it as well. For what Nozick's case seems capable of altering are our genuinely intrinsic desires, not simply our desires to do things as ways of doing what counts as good.

A final problem with Analytic Humean accounts of the influence of reflective-equilibrium thinking in terms of the above *de dicto* desires is that they seem to account for the way in which we use our reflective equilibrium methods to try to settle the whole of the basic question of what to do in a given circumstance, not just how a single consideration or a few different considerations bear on it. On the account in terms of a *de*

dicto desire to obey the dictates of reflective equilibrium methods, settling what reflective equilibrium methods prescribe would only settle what would satisfy one motive among many. But Analytic Humeans maintains that we have reason to satisfy each of our intrinsic motives. As such, this account of how we respond to reflective equilibrium methods commits them to the view that settling what reflective equilibrium methods prescribe settles only one question among many that bear on what to do.

Similarly, since we do use reflective equilibrium methods to figure out what is morally wrong and what is good, it might be plausible that we would need them to figure out what would satisfy *de dicto* desires for or against such things. But reflective equilibrium methods could not be used to determine what will satisfy many (if almost all of the rest of) our intrinsic motives, such as those to avoid our own pain, to prevent the suffering of our loved ones, and so on. So the Humean account of how we respond to reflective equilibrium methods in terms of our using them to figure out what will satisfy our *de dicto* desires to avoid wrong and promote good (and perhaps a few other *de dicto* desires and aversions regarding ethical matters) would again be committed to the view that reflective equilibrium methods can settle only how certain considerations bear on what to do rather than the entirety of the basic question of what to do.

3.3. Fitting Attitudes and Rational Ends

We have thus seen that, intuitively, what we have reason to do is not what will bring about whatever ends we are motivated to pursue, but rather what will bring about the ends that we should pursue, whether or not we are motivated to pursue them. The Analytic Humean's objection to this idea was that we have confused normative reasons for action with something else; we have reason to do something only if we could correctly reason our way to doing it, but practical reasoning is exclusively instrumental. We have seen, however, that that our practical reasoning includes a kind of basic inquiry into what ends to pursue that proceeds via reflective equilibrium methods. This kind of reasoning is able to guide the direct revision of our intrinsic motives in a way that could not be explained by the Analytic Humean as a species of instrumental reasoning. So

while the Analytic Humean is right about the conceptual connection between practical reasons and practical reasoning, she is wrong that it can ultimately save her view about what it is to judge that someone has normative reason to do something. The common sense view about practical judgments seems to stand vindicated: to judge that someone has reason to do something is to judge that it would serve the ends that she should be pursuing, whether or not she is actually motivated to pursue those ends.

What is it, though, to think that an end is worth pursuing? If performing (or omitting) some action or bringing about some state of affairs is something at which we should aim as an end, or at which we should aim independent of its further consequences, then clearly we *ought to* be motivated to perform it or bring it about as such. That is, if something is worth doing or bringing about as an end, then we have reason to be intrinsically motivated to do it or bring it about. In fact, I think that this tells us what it is for something to be an end that we should pursue – for it to be the case that we have reason to perform an act or bring about an outcome as an end in itself is just for it to be the case that we should be intrinsically motivated to do it or bring it about.

It might appear, however, that not all reasons to have intrinsic motives translate into reasons to pursue their objects as ends. Suppose, for instance, that an extremely powerful evil demon who can read your mind tells you that he will soon begin checking up on you to see if you have an intrinsic desire to kill your best friend. If you have this desire he will let your friend alone, but if you lack this desire he will kill your friend himself. This, he tells you, will go on for the rest of your lives until he either detects that you lack an intrinsic desire to kill your friend and kills him, or one of you dies of other (hopefully natural) causes. You know that you have no intrinsic desire to kill your friend, but luckily the demon shows you where you can get a program of conditioning and brainwashing that is guaranteed to produce an intrinsic desire to kill him. You reason that if you take the brainwashing course and get the desire to kill your friend, you will pose some danger to him, but chances are decent that you will be able to control yourself, and the alternative is his certain death.

In this scenario, there seems to be a sense in which you have good reason to have an intrinsic desire to kill your friend – having that desire will, after all, save him from certain annihilation. But surely this does not seem to mean that you have reason to kill

your friend as an end in itself. Indeed, you have this kind of reason to have the intrinsic desire to kill your friend even if you continue to lack the desire. Suppose that you have not yet gotten around to having the course of brainwashing and currently lack the intrinsic desire to kill your friend, but luckily the demon has not yet gotten around to checking up on you either. You then have reason (in the relevant sense) to have the intrinsic desire to kill your friend, but surely even Humeans will agree that you don't now have any reason to kill him as an end in itself.

But there seems to be an important difference between the sense in which you should have an intrinsic desire to kill your friend in the foregoing case and the sense in which you should be intrinsically motivated to pursue the ends that you have reason to pursue. Recall the distinction between on the one hand thinking an attitude fitting or warranted and on the other hand thinking it supported by what we might call pragmatic or strategic reasons – considerations that make it the case that we should do what we can to have the attitude without contributing to its warrant. The sense in which one should be intrinsically motivated to pursue what one ought to pursue as an end in itself is that such intrinsic motives are fitting or warranted. As we saw in section 3.2., when as a result of reflective equilibrium methods an agent judges that she should avoid harming non-human animals for their own sakes, she makes a judgment that it is *fitting* for her to be intrinsically averse to harming them, which, like any assessment of warrant or fittingness, has a direct propensity to cause her to have this intrinsic aversion.

But the sense in which the fact that the demon will kill your friend unless you intrinsically desire to kill him is a reason to have the intrinsic desire is that of a pragmatic or strategic reason to get yourself to have the desire. The fact that the demon has made this threat in no way contributes to the warrant, fittingness, or appropriateness of intrinsically desiring to kill your friend – it merely makes it a very good idea to get yourself to have this intrinsic desire. It is analogous to the kind of pragmatic, non-epistemic reason you have to disbelieve a theory on the grounds that the truth of the theory would make you depressed. Judgments about these kinds of pragmatic reasons for attitudes lack a direct propensity to cause you to have them. In order for your judgment that you shouldn't believe the theory on account of its depressing nature to cause you to disbelieve it you must do something to bring about your disbelief. In the same way your

judgment that you should intrinsically desire to harm your friend on the grounds that your having this desire will save his life cannot cause you to have the desire without your doing something to get yourself to have it. The demon in his wisdom understood this well, and provided the means by which you could actively bring the desire about by paying a visit to the brainwashing center (in order that you might actually respond to his threat).

With this distinction in mind, my proposal is that we can understand what it is for an end to be worth pursuing in terms of the fittingness of being intrinsically motivated to pursue it. In particular, I think that we can analyze the judgment that an end is worth pursuing as a judgment that it is fitting to be intrinsically motivated to pursue it. To think that the performance (or omission) of an act or the bringing about of a state of affairs is something at which we should aim as an ultimate end, independent of its further consequences, is to think that it is fitting to be intrinsically motivated to perform (or omit) it or bring it about. If this is right, then the following principle will be a conceptual truth about the relationship between fitting attitudes and what we should aim at in action:

> **Rational Ends Principle [REP]:**
> Let *X* be any action and let *S* be any state of affairs. Doing (or omitting) *X* or bringing about *S* is something at which we should aim as an end in itself independent of its further consequences iff it is fitting to be intrinsically motivated to do (or omit) *X* or to bring *S* about.

The analysis of judgments about what is worth pursuing in terms of judgments about what it is fitting to be motivated to pursue enables us to draw upon the understanding we developed in Chapter 2 of judgments about the latter to understand the former. As we argued there, judging a valenced attitude fitting should be understood as a basic kind of normative assessment akin to judgments about epistemic reasons for belief. Judging a valenced attitude fitting is distinguished by its direct propensity to cause one to have the attitude and our accessing its fittingness by means of reflective equilibrium methods that seek to best unify our non-debunked intuitions about fittingness. By

analyzing ethical judgments in terms of these fittingness assessments, we were able to explain how the epistemology of ethics parallels the epistemology of epistemology and how ethical judgments guide the conative components of valenced attitudes in the same way that judgments about epistemic reasons for belief guide beliefs.

But what we have seen in the case of basic deliberation about which ends are worth pursuing are the exact same causal and epistemic features that we saw needed to be explained in the case of ethical judgments. Our access to whether or not something is worth doing or bringing about as an ultimate end is constituted by means of reflective equilibrium methods that solicit our intuitions about what to pursue, explore which of these intuitions can be debunked, and seek the best unification of those that seem able to stand up to scrutiny. Our judgments that we should pursue or avoid doing something as an ultimate end have a direct propensity to cause us to be intrinsically motivated or averse to doing it. Since these are the epistemic and causal role descriptions of judgments about which intrinsic motivations and aversions are fitting, we can explain these features of thinking about which ends are worth pursuing by identifying it as such.

3.4. Ethics and Rational Ends

Indeed, we would seem to be able to give a unified explanation of the epistemology and causal powers of ethical judgments and judgments about which ends to pursue by identifying them both with judgments that certain motivational states are fitting. But in fact we can go further along these lines. Ethical judgments, we have argued, can be analyzed in terms of the fittingness of valenced attitudes - states like emotions and desires - which quite prominently involve motivations to do certain things. Because valenced attitudes themselves involve motivations, judgments about their fittingness actually entail judgments about the fittingness of these motivations. Allow me to explain this in more detail.

It is part of the very idea of various of our valenced attitudes that they involve certain motivations. It is part of our concept of a desire for a state of affairs that it involves motivation to bring about that state, part of our concept of feeling obligated to

do something that it involves being motivated to do it, and so on. Now to think a state warranted but to deny that its essential components are warranted would seem to be a kind of incoherent state of mind. It seems to be a matter of conceptual necessity that reasons to be in a psychic state are reasons to be reasons to be in all that the state essentially amounts to. If this is right, then the following principle is a conceptual truth about the relationship between the fittingness of a mental state and the fittingness of its component parts:

> **Warrant composition principle [WCP]**:
> Let ϕ be a psychic state that involves ψ as an essential component. If it is fitting to be in ϕ, then it is fitting to be in ψ.

The notion here of an essential component of an attitude is that on which we relied in Chapter 2. This was the sense of 'essential component' in which all parties were agreed that emotions and desires involve as essential components things like affect, motivation, and focused attention, and it was this sense in which I contended against judgmentalists that emotions and desires do not ethical judgments as essential components. To be an essential component of an attitude in this sense is different from being a necessary precondition for having that attitude. Being alive or having a mental life are presumably necessary preconditions for having attitudes like emotions and desires, but they are not essential components of them. To be an essential component of an attitude is to be an element that composes the attitude and forms part of what it is to have the attitude. Being alive is neither a constituent part of the attitude nor part of what it is to have that attitude.

An attitude here should also be distinguished from its intentional content. A desire that Judy have a popsicle presumably has as its content the proposition *Judy has a popsicle*. On a structured view of propositions, this proposition itself has essential components, such as Judy herself and the property of having a popsicle[99] But these essential components of the desire's content are distinct from the essential components of

[99] On an unstructured view of propositions as sets of possible worlds, it may involve as components its proper subsets, like the set of worlds in which Judy gets a popsicle and Judy wears a hat.

the desire itself.[100] If I had to illustrate all of this with a metaphor, an attitude is like a wall, its essential components are like the building blocks out of which the wall is constructed, being alive is like the solid ground on which the wall is built, and the attitude's content (and its components) are like the wall's shape (and its components). The warrant composition principle asserts only that the fittingness of the attitude entails the fittingness of its essential components or "building blocks". It in no way asserts that the attitude's fittingness entails the fittingness of its necessary conditions like being alive, its content like the proposition *Judy has a popsicle*, or the components of its contents like Judy and the property *having a popsicle*. (This is a good thing, because talk of the fittingness of these other kinds of things would seem to involve a category mistake.)

The Warrant Composition principle thus entails that if a valenced attitude is fitting, the motivations it involves are also fitting. Since desiring an outcome involves motivation to bring it about, the fittingness of the desire entails the fittingness of motivation to bring about the outcome. Since feeling obligated not to do something involves motivation not to do it, the fittingness of the feeling of obligation entails the fittingness of motivation to avoid performing the act.

Now as we have seen, we can desire certain states of affairs intrinsically, or independent of their further consequences, while we can desire others only on account of what else they bring about. Similar remarks go for other pro-attitudes towards states of affairs like hoping, wishing, or being glad that they obtain; we can have these attitudes towards states of affairs either intrinsically or on account of their further consequences too. In the same way, we can have intrinsic feelings of obligation not to do certain things – you can, for instance, feel prospective guilt-tinged aversion to harming someone merely because it will harm her and independently of what else it will bring about. There are of course other things that we can feel non-intrinsically obligated not to do, or obligated not to do on account of what else it will bring about – this, presumably, is how we feel about firing weapons at beings we feel intrinsically obligated not to harm.

[100] I am grateful to Peter Railton, Brad Skow, and Rae Langton for helpful discussion about how the essential componency relation in which I am interested differs from these other kinds of things.

If judging a state of affairs good is judging that the state befits pro-attitudes like desire, then it would seem that judging a state of affairs intrinsically good is judging it to befit intrinsic pro-attitudes like intrinsic desire. The judgment that a state is intrinsically good has a propensity to cause, not just any desires towards it, but intrinsic desires towards it. Similarly, you will tend to arrive at this judgment in normative inquiry either by inferring it rather directly from immediate intuitions that the state or states of its kind are worth wanting for their own sakes, or by thinking that this best explains or unifies some other (un-debunked) intuitions. You will also tend to draw on the judgment's content in your further use of reflective equilibrium methods to figure out what you should want, for instance as something that theories about what's good should capture. But this is just the role description of judging it fitting to intrinsically desire the state. In the same way, if judging an act morally wrong is judging that it is fitting to feel obligated not to perform it, then it seems that judging an act intrinsically wrong involves judging that it is fitting to feel intrinsically obligated not to perform it. Judging an act intrinsically wrong has a propensity to cause you to feel intrinsically obligated not to perform it; to feel prospective guilt-tinged aversion towards it quite independently of its consequences. In normative inquiry you will tend to arrive at such a judgment by inferring it from supporting intuitions or those it seems to explain, and you will tend to use it in your further use of reflective equilibrium methods, for instance as something you draw upon to in evaluating the plausibility of debunking arguments.[101]

This supports the following fitting attitude analyses of judgments about intrinsic goodness and wrongness:

> **Fitting Attitude Analysis of Intrinsically Good (and Better) States of Affairs:** To judge that state of affairs *S* is intrinsically good is to judge that it is fitting for us to have an intrinsic pro-attitude towards *S* (e.g. to intrinsically desire, wish, or be glad that *S* obtains). To judge that state of affairs *S* is intrinsically better than

[101] Thus, if you believe that breaking promises is intrinsically wrong, you might be less inclined to believe that morality is (direct) consequentialist in character and thus that organ transplant and trolley intuitions are due to crass confusions. On the other hand, if you believe that race is morally irrelevant, you may be much more inclined to suspect that intuitions about the relevance of things that look an awful lot like race (qua biological category quite independent of their members' psychology, etc.), such as sex and species membership, are due to rather crass confusions.

state of affairs S* is to judge that it is fitting for us to have an intrinsic preference for *S* over *S**.[102]

Fitting Attitude Analysis of Intrinsic Moral Wrongness:
To judge that agent *A*'s act of ϕ-ing is morally wrong is to judge that it is fitting for *A* to feel intrinsically obligated not to ϕ (or equivalently: to judge that it is fitting for *A* to feel intrinsic prospective guilt-tinged aversion towards ϕ-ing).

Now these analyses, together with the Warrant Composition Principle and the Rational Ends Principle, actually entail that the production of good states of affairs and the avoidance of morally wrong acts are rational ultimate ends. By the above analysis of intrinsic goodness, it is a conceptual truth that if a state of affairs is good, we have reason to have an intrinsic pro-attitude towards it. Since these pro-attiutdes involve intrinsic motivations to do what we can to bring about the states of affairs towards which we have them, it follows by the Warrant Composition Principle that if it is fitting to have such pro-attitudes towards a state of affairs, it is fitting for us to be intrinsically motivated to bring it about. But by the Rational Ends Principle, if it is fitting to be intrinsically motivated to bring the state of affairs, then bringing it about is something we should aim at as an end in itself. Similarly, by the above analysis of intrinsic wrongness, it is a conceptual truth that if it would be wrong for us to do something, it is fitting for us to feel intrinsically obligated not to perform it. Since this feeling of obligation involves intrinsic motivation not to perform the act, it follows by the Warrant Composition Principle that if it is fitting to have such feelings of obligation not to perform the act, it is fitting for us to be intrinsically motivated not to perform it. But by the Rational Ends Principle, if it is fitting to be intrinsically motivated not to perform the act, then omitting the act is something we should aim at doing as an end in itself.

This, I think, is rather important. For as we observed in Chapter 1, some people have suspected that we only have reason to bring about what is good if it serves some

[102] Where an intrinsic preference for S over S* is a preference for S over S* independent of what else these states constitute or bring about.

other end, like our own self interests. But if having reason to pursue something as an end is a matter of its being fitting to be motivated to do so, and the intrinsic goodness of an outcome consists in the fittingness of having a kind of intrinsic motivation to bring it about, we have reason to bring about intrinsically good outcomes whether or not they serve some ulterior purpose. If we have reason to pursue our own interests as an ultimate end, this is because it is fitting for us to be motivated to pursue whatever constitutes these interests. Thus, our having reason to preserve our own health and avoid our own pain as ultimate ends is a matter of its being fitting for us to intrinsically desire our own health and to be intrinsically averse to our own pain. But if something that is not particularly good for *us* really is intrinsically good – if it really is fitting, for instance, to intrinsically desire that children in developing countries get enough to eat – then the exact same thing that makes it the case that we have reason to pursue our own interests as ends makes it the case that we have reason pursue the feeding of the children in developing countries as an end. Of course, whether it really is fitting to want the children to have enough to eat is a substantive normative question – but so too is the question of whether it really is fitting to want to be healthy and to want to avoid one's own pain. I take the answers to all of these questions are equally obvious and equally well supported by evidence from our reflective equilibrium methods.

Similarly, some have suspected that we only have reason to avoid doing what's wrong if it serves some further purpose, like furthering our own interests, or making the world a better place, or saving us from hypocrisy, or what have you. But if having reason to avoid doing something as an end in itself is a matter of its being fitting to be motivated to omit doing it, and the intrinsic wrongness of an act consists in the fittingness of having a kind of intrinsic motivation to avoid doing it, then we have reason to avoid doing what is intrinsically wrong whether or not it serves some further purpose. For as we have seen, our having reason to promote our interests and to bring about what is good as an end in itself is a matter of its being fitting to be intrinsically motivated to do these kinds of things. But if something really is intrinsically morally wrong – if it really is fitting, for instance, to feel intrinsically obligated to keep a promise or to prevent a being from coming to harm – then the same thing that makes it the case that we have reason to promote our interest and to bring about what is good as ends makes it the case that we

have reason to avoid doing these things as ends in themselves and independent of their further consequences. Again, whether it is fitting to feel intrinsically obligated not to do something is a substantive normative question, but again too there seems to be nothing more mysterious or doubtful about the fittingness of intrinsic feelings of obligation than the fittingness of intrinsic desires for what promotes one's own interests or makes the world a better place.

I believe that this explains how we have the kind of "intrinsic" or "right kind of reasons" to avoid doing what is morally wrong that people like H.A. Prichard (1912) and W.D. Falk (1948, esp. 23) were concerned with. It is important to note, however, that this explanation shows us to have genuinely intrinsic reason to avoid doing what is morally wrong *de re* rather than what is morally wrong *de dicto*. That is, according to this account we have intrinsic reason to avoid doing are *the things* that are intrinsically wrong, we do not (simply) have reason to avoid doing these things *for the further reason* that they are morally wrong. For instance, we are guaranteed that if breaking promises is intrinsically wrong, then we have reason not to break a promise simply because it is breaking a promise. For the intrinsic wrongness of promise breaking would amount to the fittingness of an intrinsic feeling of obligation not to break promises rather than an intrinsic feeling of obligation not to do whatever is *de dicto* morally wrong and an instrumental feeling of obligation not to break promises as a way of avoiding what is wrong. It is the fittingness of the genuinely intrinsic feeling of obligation that combines with the Warrant Composition Principle and the Rational Ends Principle to entail that we have reason to be intrinsically averse to breaking promises and to avoid breaking promises as an end in itself. Similar remarks hold, of course, for the way in which this pattern of explanation shows us to have reason to bring about intrinsically good outcomes *de re* rather than *de dicto*. If children's happiness is intrinsically good, then the fact that a state involves children's happiness is an ultimate reason to bring it about; it is not that we (merely) have reason to bring it about for the further reason that it will be the bringing about of something good. For what the intrinsic goodness of children's happiness entails is the fittingness of an intrinsic pro-attitude to such happiness, and thus intrinsic motivation to bring it about and reason to bring it about as an end in itself.

Another important feature of this account is that it explains why we have reason to bring about what is intrinsically good and why we have reason to avoid doing what is intrinsically wrong in a way that is independent of the truth of any particular substantive theory of what is good or wrong. For recall that the fitting attitude analyses of intrinsic goodness and wrongness, the Warrant Composition Principle, and the Rational Ends Principles appear to be conceptual truths, or truths that hold whatever substantive theory of what is good and wrong turns out to be true. In particular, this means that we have reason to avoid doing what is intrinsically wrong whether or not this conduces to the best overall state of affairs and whether or not we would be hypocrites if we did what is wrong. I believe that this undercuts attempts to argue in favor of views like consequentialism and certain forms of Kantian or contractualist deontology on the grounds that these views best explain why we have reason to be moral. This explanation of why we should be moral is of course consistent with the truth of any such substantive view. Indeed, I think that there is a good deal to be said in favor of direct consequentialism about moral wrongness, here analyzed as the view that what it is fitting feel intrinsically obligated to do is the same as what would make the world as good a place as possible. But we would not need any such identity between what is right and what makes the world better to explain why we have reason to avoid doing what is wrong. As such, an alleged ability to explain why we have reason to be moral is not among the things that can be truly said in favor of the view.

More generally, we might think that assessments of the intrinsic ethical status of acts, states, and characters can be analyzed in terms of the fittingness of intrinsic kinds of valenced attitudes towards them. Thus, to judge an act intrinsically lowly is to judge that it fitting to have an intrinsic aversion to performing it out of one's sense of honor, to judge an act intrinsically praiseworthy is to judge it to befit esteem intrinsically, or independent of its further consequences, and so on. Further, we might suspect that these other intrinsic varieties of valenced attitude involve intrinsic motivations to do various things. So an intrinsic aversion to doing something out of a sense of honor involves a motivation to avoid doing it as an end in itself, intrinsic esteem of an act involves intrinsic motivation to emulate it, etc. If this is right, then these fitting attitude analyses

of assessments of intrinsic ethical status, in conjunction with the Warrant Composition Principle and Rational Ends Principle, will guarantee that facts about intrinsic ethical status entail facts about what ends are worth pursuing. Just as a state's intrinsic goodness or an act's intrinsic wrongness entails that we have reason to bring it about or avoid doing it as an end, an act's intrinsic lowliness or the intrinsic estimability of someone's act will entail that we have reason to avoid doing it or emulate her doing it as final ends.

3.5. Fitting Attitudes and the Normative Guidance of Action

We have thus seen how accounting for judgments about which ends are worth pursuing as judgments about what it is fitting to be intrinsically motivated to do can explain the causal and epistemic features of the former, as well as the way in which ethical thought commits one to thought about what one has reason to pursue as an end in itself. But the account of what to pursue in terms of what it is fitting to be motivated to pursue has another, closely related virtue as well. This is its ability to help explain how our actions are guided by our views about what to do.

Several authors have been drawn to the idea that there is a close connection between the rational control of our motivations and the rational control of our actions, and that this tells us something about the relationship between normative reasons for one and normative reasons for the other. The guiding idea here is put well by Darwall, who tells us that "only what can be normatively guided can be subject to oughts" (Darwall 2006, 286), or "only what can be regulated by norms…can be subject to normative judgment" (Darwall 2003, 478). Along these lines, Bratman suggested that we understand the rational status of actions in terms of the rational status of the intentions that lead us to them, on the following grounds:

> A rational agent's control over her actions goes by way of her intentions. She does not separately control her present-directed intentions and her intentional actions. So there would seem to be no point in supposing that the agent escapes criticism for her present-directed intention to *A*, and yet still is subject to criticism for having failed to satisfy the relevant standards in intentionally *A*-ing (Bratman 1987, 54).

In a similar way, Scanlon (1998, 20-21) suggested that "judgment sensitive attitudes" – including beliefs, the valenced attitudes we have been discussing, and intentions – constitute "the class of things for which [normative] reasons…can sensibly be asked for or given." Scanlon notes, however, that normative reasons clearly apply to acts, and that acts are not any kind of judgment sensitive attitude. But, Scanlon argues:

> Actions are the kind of things for which normative reasons can be given only insofar as they are intentional, that is, are the expression of judgment-sensitive attitudes…A reason for doing something is almost always a reason for doing it intentionally, so "reason for action" is not to be contrasted with "reason for intention" (Scanlon 1998, 21).

What Bratman and Scanlon seem to have in mind, then, is that since our practical reasoning controls our actions by means of controlling our intentions to perform them, we can understand reasons for action as reasons for the intentions that generate them.

But how are our intentions themselves rationally controlled? Does their rational control have something to do with the rational control of our other motivational states, like the desires and emotions we have been discussing? For this we should turn to Bratman's work on the relationship between intentions and what he calls 'desires'. Desires, according to Bratman, are motivational states like those involved in valenced attitudes such as emotions and what we have been calling 'desires' in the thicker, directed-attention sense.[103] Intentions are also a kind of motivational state, in that they share with these other states the basic functional role of combining with representations of what will realize their objects to produce action. But intentions are different from other motivational states in they involve a kind of "commitment" to a plan of action that other motivations lack. Following, Bratman we might contrast a desire to have a bag of chips for lunch with an intention to do so. Even if the desire is very strong, it must be

[103] According to Bratman "Desires include a wide range of "pro-attitudes" – wanting, judging desirable, caring about, and so on" (Bratman 1987, 6). Of course, if the arguments of Chapter 2 (in particular section 2.2) are correct, we cannot regard judging something desirable as itself a pro-attitude – it can be had bloodlessly, and while it has a propensity to cause desire, it has different functional roles to play. (While for our purposes we must distinguish judgments of desirability from motivational states proper, we shall see later on that there is actually a pathway by which judging something desirable can play much the same kind of role in the formation of intentions that Bratman attributes to what he calls 'desires'. So, even if the arguments of Chapter 2 are correct, there is a sense in which Bratman is not mistaken to categorize judgments of desirability with motivational states in light of his focus on intentions and their relations to other mental states).

weighted against one's conflicting desires, like one's desire to lose weight, if it is to determine what one does. The desire might well be outweighed by other desires, or one might fail to act on it for other reasons. But one can in this way retain the desire and fail to act on it without any failure of rationality. An intention to have the bag of chips, on the other hand, will typically not need to be weighed against one's other motives to determine what one does; when one intends to have the chips one has already decided in advance to have them. One will typically try to execute one's intention, and if one retains the intention but fails to act on it, one will be guilty of a kind of irrationality. While it seems fine to retain desires to do things that one does not do, one had better jettison an intention to do something if one is no longer going to do it.

Bratman refers to this feature of intentions by calling them *conduct controlling* motivational states, and contrasts them with the *potential conduct influencing* nature of desires and other motivations that are distinct from intentions.[104] This is an important part of the way in which intentions involve commitment to a course of action in a way that other motivational states do not. But Bratman also points to a second important feature of how intentions involve commitment to a course of action, which concerns their role as relatively fixed points in deliberation about what to do. Suppose that early in the day one thinks that working over at the library would be a good thing to do with the afternoon, and that one has a corresponding desire to spend the afternoon at the library. Having this view and desire in no way tends to treat the matter of whether to go as settled or to close off deliberation about whether or not to go. But having an intention to go to the library does seem to treat the question of whether to go as settled and close off deliberation about whether to go. One can, of course, reconsider one's intention if important new information comes to light, but the intention has a tendency to resist reconsideration. A related feature of intentions is their role in structuring reasoning about how to realize their objects, and to further more specific intentions. Intending to go to the library give rise to thinking about how to get there, to intentions to take a certain means

[104] Bratman actually calls intentions 'conduct controlling pro-attitudes', but what Bratman means by 'a pro-attitude' is just what I mean by 'a motivational state', so in order to avoid confusion (and to reserve the phrase 'pro-attitude' for the "positively valenced" attitudes the fittingness of which is used to analyze ethical certain judgments, which are not intentions) I stick to my terminology of motivational states.

of transportation, to be where one need to be to take it, and so on. We might call this feature of intentions their *reason guiding* role.

As Bratman points out, intentions play an important role in the lives of cognitively limited being like ourselves who face pressures to act, but who can nevertheless see some ways into the future and reason about what to do. Intentions help keep us focused on the tasks at hand in the face of an otherwise bewildering array of options. Given our ability to look into the future but our limited ability to reason at any given point in time, intentions enable us to reason about what to do in a situation before it arrives, enabling us to do better when time comes to act than we would have done had we had to calculate the net strength of our reasons on the spot. Intentions also help us with the task of acting when we are faced with *Buridan cases*, or cases in which several of the courses of action open to us are equally desirable or equally desired.[105] In such cases, we seem to have the ability to pick one of our equally preferred options, and intentions help us stay in pursuit of the selected option in the absence of any reasons to prefer it or any other motivations to pursue it over its alternatives.

These features of intentions require them to be governed by reason in a way rather different than our other motives, but in a way that relates intimately to their governance. We have seen how intrinsic desires and aversions are governed by our assessments of them as fitting, and how these assessments constitute our thinking that we have reason to realize the objects of these motives as ultimate ends. Since the fittingness of intrinsic motivation to do something is analytically equivalent to its being something worth doing for its own sake, considerations that count in favor of having such an intrinsic motive (in the sense of contributing to its fittingness) count in favor of doing what realizes its object for its own sake. Thus, if the fact that an outcome involves children being happy counts in favor of intrinsically desiring that outcome, then this fact about the outcome counts in favor of bringing it about as an end. Similarly, if the fact that an act would constitute a breach of promise counts in favor of feeling intrinsically obligated not to perform it, then this fact about the act counts in favor of omitting it on its own account.

[105] The cases are named after Jean Buridan who (perhaps apocryphally – see Ullmann-Margalit and Morgenbesser (1977)), worried about what might happen to a donkey who had the ability to engage in practical reasoning, responded only to what he took to be the strength of his practical reasons, and was placed mid-way between two equally desirable piles of hay.

Thus, in judging an intrinsic desire or aversion fitting, we take ourselves to have reason to bring about its object quite independently of whether it brings about anything else, including the objects of our other motives. Barring recalcitrance or bloodlessness, such a judgment gives rise to the desire or aversion, which in this sense reflects our view that we have independent reason to bring about its object. But if intentions are to play their roles in keeping us focused on the task at hand and enabling us to execute decisions among equally desirable options, they cannot in this kind of way reflect our judgments about a new set of reasons to bring about their objects. Rather, we must take ourselves to have no more reason to do what we intend to do than we have to be inclined to do it by motivational states that are not intentions.

Our intentions, as Bratman argues, simply reflect the views about our reasons for action that are already reflected in our desires and aversions.[106] Suppose that you are in a boat, equidistant between two drowning children, Bugsy and Suzy. You can either row left to save Suzy or row right to save Bugsy, but you cannot save both. Your view that Suzy's survival counts in favor of rowing left is reflected in your desire to row left, and your view that Bugsy's survival counts in favor of rowing right is reflected in your desire to row right. But unless you act quickly both will drown, so you must pick. Let us say that you pick to row left and save Suzy. Your intention to row left will then reflect your view that Suzy's salvation counts in favor of rowing left – the same view that is reflected in your desire to row left. But it does not reflect your view that there was some further consideration in favor of rowing left and saving her. Even as you row left, you can agree that there really was equal reason to save both children; it was simply that you had to act, and the reasons on which you acted were the reasons to save Suzy.

Thus, if Scanlon is right that reasons for action are not to be contrasted with reasons for intention, it seems that reasons for intention cannot themselves be contrasted with reasons for desire and aversion. We can, of course, speak of "reasons for intention"

[106] Bratman often talks as though these desires and aversions were themselves (normative) reasons for action, but he does intend to remain neutral as between something like the Humean theory of practical reasons and the view that our intrinsic desires respond to our views about what's appropriate to desire or what's worth pursuing as an end in itself. He invites those of us who are non-Humeans to replace his talk of desires with talk of rational desires (on page 22 and in the form of parenthetical additions of 'ration' behind 'desire').

in a sense that covers cases of what we might call pragmatic or strategic reasons to form intentions. Kavka's (1983) Toxin Puzzle features reasons to form intentions of this kind. In this puzzle, an eccentric billionaire will pay you $1,000,000 if at midnight tonight you have an intention to drink a toxin tomorrow morning which will make you sick for a day (whether you have the intention will be detected by a completely reliable brain-scanning device). In this case, it seems that in a sense you have reason to intend to take the toxin, but it surely does not seem that you have reason to act out of the intention and actually take it - after all you are paid only for having the intention tonight; actually taking the toxin tomorrow gets you nothing but sick.

But recall again the distinction between considerations that contribute to an attitude's warrant and considerations that contribute to one's merely having pragmatic reasons to get oneself to have it. In the toxin puzzle, the fact that you will get $1,000,000 for having the intention do not contribute to its warrant or fittingness, they merely contribute to its being a good state to get yourself to have. Contrast this with your reasons to intend to, say, visit the dentist tomorrow. The fact that visiting the dentist will relieve you of your pain does not simply make having the intention to visit the dentist a good idea; it actually contributes to the warrant or fittingness of the intention. (Similarly, the fact that rowing left towards Suzy is needed to save her life contributes to the warrant or fittingness of the intention to row left. Of course, the fact that rowing right will save Bugsy equally contributes to the fittingness of an intention to row right. But one nice thing about intentions is that there is nothing wrong with forming one but not the other of two conflicting but equally warranted intentions.)

As Kavka pointed out, merely taking yourself to have pragmatic reason to get yourself to have the intention to take the toxin is insufficient on its own to get you to have it. To respond to your perceptions of these kinds of reasons you rather must do something else, like bind yourself to taking the toxin by, for instance, promising to take it or signing a legally binding agreement to give money to your least favorite political party if you don't.[107] But taking it to be fitting to intend to do something can directly cause

[107] Ralph Wedgewood has suggested the following case that might seem to be an exception. Suppose that an eccentric billionaire will give you $1,000,000 if you simply form an intention to raise your arm. Clearly, it seems, you can easily form the intention and walk away with the $1,000,000 by simply raising your arm and without the need to bind yourself, make side-bets or whatever. So it might look as though we

you to intend to do it. Taking yourself to have reason to intend to visit the dentist on account of the fact that the visit will relieve your pain can give rise to the intention to visit the dentist without your having to do anything to bring this intention about. When everything is working properly, you need not bind yourself to the action of visiting the dentist in order for you to intend to visit her in response to your perception of the need to relieve your toothache.[108] Like other judgments of fittingness or warrant, judgments to

---

have here an instance of a pragmatic reason to form an intention to which you can respond directly, or without your having to do anything to bring it about. But in this case you still must do something in order to bring it about that you have had the intention, namely raise your arm. Unlike in the Toxin Puzzle, you simply happen to be lucky that you have available to you a basic action that you know to be such that if you perform it, you will cause the intention to be brought about that will be rewarded. What you realize is that raising your arm will enable you to claim the reward.

Thus, the fact that you will be rewarded for having the intention to raise your arm, given the fact that you can bring this about by raising your arm, actually contributes to the fittingness or warrant of the intention to raise your arm with the further intention of bringing about the intention of raising your arm. It is this consideration of warrant to which you can directly respond by intending to raise your arm and doing so. Suppose, however, that we altered Wedgewood's case so that the billionaire will not pay you for having the intention to raise your arm *if you intend to raise your arm with the further intention of bringing about the intention to raise your arm by so doing*. Then there is no consideration of warrant for the intention to raise your arm to which you can respond, and you face a case just like the Toxin Puzzle, in which you cannot form the intention for which you will be rewarded without doing something else, like making a side-bet, taking a mind-altering substance, or what have you.

As we shall see in Chapter 6, the fact that we have to perform physical actions in order to bring about attitudes like intentions in response to pragmatic reasons is a contingent feature of our psychology that could have be otherwise (it is conceptually possible to have a creature who can form intentions as a kind of basic action). But what does not seem to be a contingent fact is that one's responses to what one takes to be pragmatic reasons must be preceded by a response to what one takes to be reasons of warrant or fittingness (our creature, for instance, would have first to take it to be fitting to undertake the basic action of forming whatever intentions he has pragmatic reason to form).

[108] Though doing this kind of thing can help in the presence of a certain kind of weakness of practical reasoning. You might conclude that you really ought to go to the dentist, but find yourself unable to intend to do so (unable, for instance, to really treat the matter as settled and begin thinking seriously about how to get there, to be unable to be in the kind of state that really will tend control your conduct tomorrow in the direction of going there, and so on). Faced with this, you might find yourself able to form the intention after making a side-bet with a friend that you will go, which will pay you some money if you do see the dentist after all. This kind of weakness of practical reasoning is akin to typical cases of (what I would think we should call) weakness of will in that you find yourself insufficiently motivated to do as you think you ought, except that where weakness of will involves a failure to do as you intend, this kind of weakness involves is a failure to even intend to do what you think you ought to do.

(It seems, however, that some authors seem to describe the above weakness of practical reasoning as a case of weakness of will, and some seem to think that weakness of will in the sense of engaging in motivated behavior contrary to one's intentions is impossible (indeed, some like Hare (1952) have held that both this *and* failing to do what one judges one ought to do is impossible). I think, however, that the distinction forced upon us by Bratman between intentions and other motivational states - where intentions are not simply our strongest motivations - requires that we recognize this possibility. Conduct controlling though they may be, intentions may lose the battle of controlling conduct to sufficiently strong desires and aversions. I take this to be the natural way of describing what happens when one intends not to eat chips but one is overcome by a desire to eat them. Since our views about what to intend have a propensity to influence our intentions, this kind of thing is closely correlated with the above weakness of practical reasoning in controlling our intentions. But the two are still distinct. One might judge a certain intention

the effect that certain intentions are fitting have a propensity to cause them directly, without your trying to bring them about, while judgments about pragmatic reasons for intentions, like other such pragmatic reasons, lack this propensity.

Thus, while we can have a pragmatic reason to intend to do *X* without having reason to do *X*, reasons to intend to do *X* that contribute to the fittingness of an intention to do *X* seem to be none other than reasons to do *X*. Clearly it was the latter kind of considerations that contribute to an intention's fittingness that Scanlon had in mind when he spoke of the identity of reasons for action with reasons for intention. What we have learned from Bratman's observations about the dependence of reasons for intention on reasons for desire and aversion is thus the following. The considerations that contribute to the fittingness of desires or motivations to do *X* are *ipso facto* the exact same considerations that contribute to the fittingness of an intention to do *X*. Our intentions themselves are guided by these fittingness assessments, and this guidance of intention subsequently controls the actions we perform intentionally. So if the normative guidance of action by means of the normative guidance of intention enables us to understand reasons for action as reasons for intention, the normative guidance of intention by means of the normative guidance of motivations other than intentions equally enables us to understand reasons to be motivated to do something as reasons to intend to do it, and thus (by the foregoing) as reasons to do it.

Let us take a closer look at exactly how our assessments of what is worth wanting or pursuing influence our intentions and actions. As we have seen, judgments that it is fitting to be intrinsically motivated to do something tend to give rise to the corresponding motivation. And as Bratman has observed, we often form our intentions as a result of our views about what will bring about the ends we desire. A natural picture of how fittingness assessments influence motivation, intention, and action, would thus be the following. We first take certain intrinsic motives to be fitting, and as a result tend to have

---

irrational, fail to abandon it, but also (perhaps luckily) be overcome by a desire to do otherwise. Thus, a prospective bank robber might intend to rob a bank, but come to judge that his reasons not to take others' money are actually stronger than his reasons to rob the bank. Unfortunately, this realization might well be insufficient to cause him to abandon his plan – he still finds himself earnestly plotting with his co-conspirators, renting the getaway car, etc., and really is in a state that will control his bank-robbing conduct to completion. But just before going in to rob the bank he might be overcome with terror at the thought of getting caught and run away in a panic).

these motives (in a parallel way, our views about which beliefs are warranted in light of our evidence give rise to certain "means-ends beliefs," or beliefs about what will bring about the objects of our motives). Our intrinsic motives then combine with our views about what will bring about their objects to give rise to a set of preferences to do various things. We can think of a preference to do something as one's "net" motivation to do it given one's beliefs about the extent to which it will realize the objects of one's various intrinsic motives. One's various preferences over actions determine a set of "most preferred actions" – a set of things we can do in our circumstances that we take to serve our intrinsic motives equally well, and better than anything else we could do. We then pick one among our set of most preferred actions, which gives rise to an intention to perform it (or, if there is a unique most preferred action, we just form an intention to perform it), and this intention causes us to perform the action.[109] This picture of how practical reasoning regulates our motivations, intentions, and actions is depicted below in figure 5.

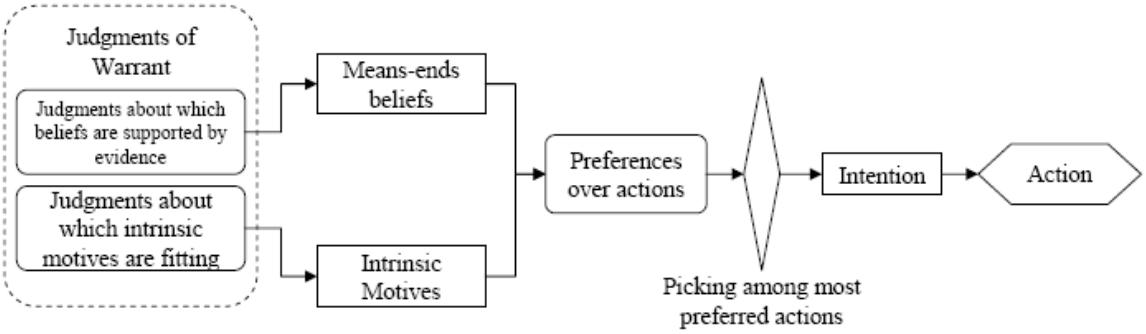

Figure 5. How practical reasoning guides motivation, intention, and action

Thus, when you are equidistant between Suzy and Bugsy on your rowboat, your view that you owe the same to each of them and that each is equally worth saving gives rise to an intrinsic motivation to help Suzy and an equally strong intrinsic motivation to help Bugsy. The role of your judgment about the fittingness of these equally strong

[109] As Bratman discusses, an important part of how motives and beliefs give rise to intentions is the use of prior intentions as "screens" on admissible options. This is most important for understanding how we operate with our cognitive limitations and how intentions develop over time. But since my focus is primarily on objective reasons for action at a given time, we can largely neglect the screening process, or understood it to be folded into the process of picking among equally preferred options.

motives is important; if you were a sexist who thought men were more important than women you might well have a stronger intrinsic motive to save Bugsy, if you were a racist and Bugsy is not of your race you might have a stronger intrinsic motive to save Suzy, and so on. These intrinsic motivations, in conjunction with your views about what will happen if you do various things give rise to a set of equally preferred actions consisting of rowing left in order to save Suzy and rowing right in order to save Bugsy (where your preferences to do these things are stronger than other preferences you might have, like staying put out of laziness or what have you). You then pick among your most preferred actions – as it happened above you pick to row left, you form an intention to perform that action, and finally you execute that action.

This picture of how practical reasoning guides our motivations and actions seems plausible for how things work much of the time, when everything is going smoothly. In it our assessments of which intrinsic motivations are fitting largely play a "background" role in explaining what intrinsic motives we have. After these fittingness assessments cause us to have the intrinsic motives we do, "instrumental reasoning" takes over and guides us pretty much to the point of action on the steam of our intrinsic motives and means-ends beliefs. When this is how our practical reasoning guides us, our philosophical views about what is really worth pursuing need not constrain us to act against our desires, emotions, and inclinations, for our intrinsic conations are perfectly in tune with our views about what they should be. This would seem to describe what we might think of as the way in which a sage or maximally virtuous person is guided by her practical reasoning, in which she need not exert intentional control over her actions in opposition to her intrinsic inclinations.

But it is important to note that even when our views about what to seek as an end function in this way as background controllers of our intrinsic motives, the purported normative reasons for which we intend and act are still those that we take to contribute to the fittingness of our motivations and preferences. As we have seen, we do not take ourselves to have a distinct set of reasons to intend to do things above and beyond those that we take ourselves to have to desire the things that our intentions realize. When we take an intrinsic motivation to save Suzy to be fitting, we take the fact that an act will save Suzy to count in favor of a preference for performing that act. Thus, we will take

the fact that rowing left will help save Suzy to be a reason in favor of a preference for rowing left. Our intention to save Suzy (rather than Bugsy) is formed by picking among options we take ourselves to have equal reason to perform, and the reasons we take to support our intention to row left and our action of rowing left are the very same considerations that counted in favor of the preference to row left – namely the fact that it will help save Suzy. These views about what justifies your motivations and preferences, quite apart from whether you actually have them, account for your intentions and actions even when your motives and desires are in line with your views about what to desire and prefer.

Talk of an agent's doing things for reasons can extend beyond what he takes to be normative reasons. Thus, to adapt an example of Quinn's (1993), I might find myself with a recalcitrant intrinsic desire to turn on radios – a desire that I regard as unwarranted, but that has a tendency to cause me to do things nonetheless. But if asked why I am walking towards the next room where I believe there to be a radio, you still might reply that I am doing it for a reason – namely to get to a radio and turn it on – even though you know that I am doing this out of what I regard as a compulsion and in no way think that turning on radios is something I have reason to do for its own sake.

Yet we can also speak of an agent's doing things for what she took to be normative reasons, or considerations that counted in favor of doing it. We are able to give such explanations of what an agent does in terms of her views about why she should do it because her normative views play a distinctive causal role in shaping her conduct. When we say that I am walking to the next room in order to get to a radio, we are giving the following kind of explanation of my behavior: my belief that I can get to a radio by walking to the next room combined with my desire to turn on radios to produce the behavior. Since I do not take the fact that I can get to a radio by going to the next room to count in favor of going to the next room, any such view about my normative reasons is of course irrelevant to the explanation of what I do. But when we say that you are rowing in the direction of Suzy because you take Suzy's salvation to count in favor of the rowing, we mean more than that your belief that by rowing left you can save Suzy combined with your motivation to save Suzy to cause your rowing. We mean this, but

we also mean that this happened *because* you took Suzy's salvation to count in favor of rowing left.

Part of the explanation of how your normative views accounted for the way in which your motives combined with your beliefs to produce your behavior is that these normative views explain what you were intrinsically motivated to do in the first place. The view that saving Suzy was a goal at which to aim explains why you had an intrinsic motivation to save her, which combined with your beliefs to give rise to your preference to save her and thence your intention to save her and your action of saving her. But this is only part of the story. For it seems that even if your motivations and preferences had not responded in such a nice way to your normative views, you would not have been "stuck" with these motivations and preferences. Your taking yourself to have the reasons you did explains why you "went with" your motivations and preferences, and why you did not try to oppose them and do something else.

To appreciate how our normative views can guide action in opposition to our motivations and preferences, we need to consider how practical reasoning can guide us when something goes wrong. As we saw vividly in Chapter 2, our valenced attitudes do not always conform to our views about what they should be. Our valenced attitudes can be recalcitrant in the face of our views that they are unfitting, and our ethical judgments that we really ought to feel a certain way can be bloodless. Thus, a depressed or apathetic person might genuinely judge children's happiness to be intrinsically good, but find himself with no real desire for their happiness, and you might think that it would be morally wrong to steal from someone you hate, but find yourself with no real feeling of obligation not to steal from him. The failure of our attitudes to respond to our judgments of their fittingness saps us of the motivational force that these attitudes would supply. But as we also saw in Chapter 2, we are not altogether powerless in the face of recalcitrant attitudes. One of the things that helped us see how ethical judgments could be bloodless was our observing that judging an attitude fitting can actually play a role in causing much the same behavior that would have been caused by the attitude itself. Thus judging children's happiness to be good or to befit desire can cause you to try to bring about their happiness even when you lack the desire to, and judging stealing from the

hated to be wrong or to befit prospective guilt-tinged aversion can cause you not to rob them even when you lack guilt-tinged aversion towards such theft.

Thus, it seems that judgments that a motive is fitting can guide our actions even when they fail to guide our motivations. How is this accomplished? If you judge that children's happiness is good, but in a state of depression you fail to desire to bring it about, you can, it seems, "suck it up." You can realize that children's happiness is something that you really should want for its own sake, that it really is intrinsically desirable, and in light of this realization you can steel yourself to act so as to bring it about. This process of steeling oneself and acting despite one's inclination to do nothing requires voluntary effort. It requires the formation of an intention to do what one actually lacks a desire to do. Similarly, if you judge that it would be wrong to rob or assault someone, but given your deep and abiding hatred for him you are strongly inclined to do these things and don't actually feel obligated not to do them, you can "suck it up" in much the same way. You can realize that you really should be intrinsically averse to such assault or theft, that no matter how much you might hate him you really do owe it to this person not to take his stuff or beat him, and steel yourself to refrain from doing these things despite your strong inclinations to do them. This too requires voluntary effort, which is exercised by an intention to avoid doing what you actually feel no aversion to doing.

It seems, then, that our views about what it is fitting to be intrinsically motivated to do, or what we should be aiming at as ends, are able to cause us to form intentions to do what serves these ends even when they fail to generate the corresponding intrinsic motivations. In such cases, our intentions seem to be shaped by our means-ends beliefs, not about what will satisfy our actual preferences over acts, but rather about what will satisfy the preferences over acts that we *should* have. Suppose that you are again in your rowboat equidistant between Suzy and Bugsy, and you judge that you have powerful and equally strong intrinsic reasons to save both of them. But this time let us suppose that you are so deeply depressed that you barely feel anything; in particular you feel no inclination to save either of them, no intrinsic guilt-tinged aversion to failing to save them, no aversion to their dying – in general no kind of intrinsic motivation to save them whatsoever. Still, as we have seen, you can suck it up. Although you do not desire or

feel obligated to save either of them, you can realize that you really should. You can reason that because saving Suzy is a really important goal, and rowing left can save her, rowing left is a well supported or choiceworthy thing to do. Symmetrically, you can reason that because saving Bugsy is just as important a goal, and rowing right can save him, rowing right equally well supported and choiceworthy. Although you feel no inclination to take either of these means to either of these ends, you can see that the ends that would be accomplished by these means are more important than anything that could be accomplished by doing something else – than anything, for instance, that would be accomplished by standing still where you are. So you can see that given the weighty and equal importance of the ends you really ought to pick either to row left or to row right. And this, it seems, can cause you to pick to row in one of these directions and to row in the direction that you pick in spite of apathy to rowing at all.

There thus seem to be two ways in which our practical reasoning can guide our motives, intentions, and actions. In what (we may hope) is the ordinary case, where things are functioning smoothly, our judgments about the fittingness of intrinsic motives gives rise to these motives, these motives together with our means-ends beliefs give rise to a set of acts we are most inclined to perform, we pick one of these acts and intend to perform it, and we perform it. But when our intrinsic motives fail to yield to our assessments of their fittingness, practical reasoning seems to have a kind of "backup plan," or alternative route by which it can guide us. Faced with recalcitrant intrinsic motives, we can "suck it up," and directly combine our judgments about what would be fitting to be motivated to pursue with our means-ends beliefs to yield views about how much we should be inclined to perform various acts in light of the importance of the ends they would serve. These views about rational preferences over acts determine what we take to be a set of "most choiceworthy actions" - a set of actions that we take to serve the ends at which we should be aiming equally well and better than anything else. We then pick one among this set of acts that we take to be most choiceworthy, which gives rise to an intention to perform it (or again, if we take there to be a unique most choiceworthy action we just form an intention to perform that one), and this intention causes us to perform the action in spite of our lacking any actual motivation to perform it prior to our formation of the intention. Figure 6 illustrates this picture of how practical reasoning

guides us both in the ordinary cases of smooth functioning and in cases in which our motives are recalcitrant and we must "suck it up" to respond to our views about what to do by exerting intentional control to act contrary to our non-intentional motivations.

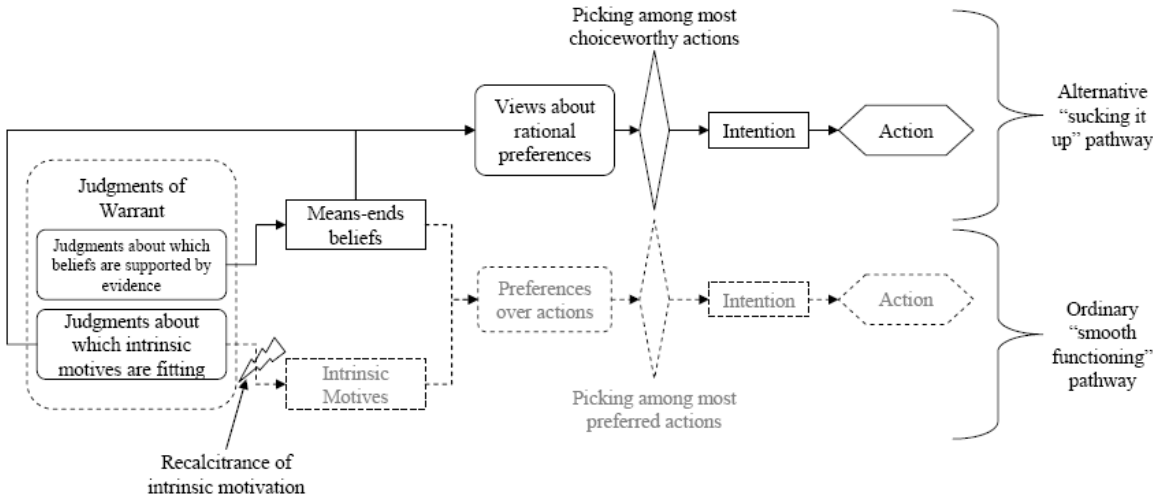

Figure 5. Guidance by practical reasoning: the smooth functioning and "sucking it up" pathways

When our intrinsic motives fail us and our practical reasoning must guide us by means of our sucking it up, our normative views about what is really worth pursuing are very much at the front of our minds. Here our instrumental reasoning is very clearly in the service of bringing about ends we think worth achieving instead of ends we are actually motivated to achieve. Since in cases of this kind we fail to be moved to pursue the ends we think we should, we are guilty of a kind of practical failing that makes us less than maximally sage-like or virtuous. But of course if we should in fact manifest this kind of failing, there is an important kind of second-best virtue involved in the disciplined control of what we do in light of our views about what ought to be done. I shall decline to speculate about how someone like Kant might have been attracted exclusively to this sort of virtue by failing to recognize that our intrinsic motives or inclinations are themselves guided by the assessments of which ends it is fitting to be moved to pursue, which stand at the foundation of all practical reasoning (as I shall decline to speculate as to whether such persons confused the specifically moral motive of feeling obligated with the genuinely domain general motivational state of intention). I

shall simply say that if anyone were to exhibit such a pattern of attraction, she would be right to recognize the controlling influence of our views about what's worth pursuing, rather than what we happen to be motivated to pursue, when we are guided by our views about what to do. She would simply be wrong to think that what we happen to be motivated to pursue is always insensitive to our views about what's worth pursuing, or that our practical reasoning cannot work *through* our intrinsic inclinations when they *are* in fact responsive to our assessments of their fittingness.

We have thus seen that whether or not our motives respond smoothly to our fittingness assessments, our actions are normatively guided by means of the normative guidance of our motivations. Most proximally, our actions are normatively guided by our views about which intentions are fitting, but these in turn are normatively guided by our views about what it is fitting to prefer to do most, which are in their turn normatively guided by our views about what it is fitting to be intrinsically motivated to do, or motivated to do as an end in itself. The normative guidance of our intentions by our views about what it is fitting to be moved to pursue and what it is fitting to prefer to do is perhaps most vivid when our motives and preferences are themselves recalcitrant, for in this case our views about fitting motives combine directly with our means-ends beliefs to generate intentions. But this guidance is no less present when our motives and preferences do respond to our fittingness assessments and play the role of combining with means-ends beliefs in the formation of intentions. For our views that these motives and preferences are fitting explain why we go effortlessly with these motives and preferences in forming our intentions, and why we do not rather spurn these attitudes and form our intentions on the basis of what we think they should be.

Now we have on hand a way to identify judgments about what to do with judgments about what will realize the objects of fitting motivational states. We saw that identifying the judgment that something is worth doing as an ultimate end with the judgment that it is fitting to be intrinsically motivated to do it could explain the epistemic features of the former as well as its causal influence on our intrinsic motivations. We also saw that common sense has it that to judge that we have reason to do something as a means is to judge that our doing it will help bring about something that it is worth

bringing about as an end. In the objective sense of 'fitting', or fitting in light of the facts of our case rather than our evidence about them (which we have been assuming throughout), what it is fitting to be motivated to do as a means is just what will realize the objects of what it is fitting be motivated to do as an end. Thus an intention or preference, like one to row left, is fitting just to the extent that it realizes the object of a fitting intrinsic motive, like a motivation to save Suzy. So the judgment that we have reason to do something as a means can be identified with the judgment that it is fitting to be motivated to do it as a means. In this way we can equate judging that we have reason to do something, whether as an end or as a means, with the judgment that it is fitting to be motivated to do it, whether as an end or as a means.

Since our actions are guided by our assessments of the fittingness of the motives that lead us to them, we can explain how our actions are guided by our judgments about what to do by identifying them with our judgments about which motives are fitting in the way just described. For then we can see how the reasoning that guides our intrinsic motivations, preferences, intentions, and actions is all practical reasoning, or thinking about what to do. As Bratman and Scanlon suggested, we should not distinguish normative reasons for the motivational states that lead us to action (normative reasons, that is, that contribute to the *fittingness* of these states) from normative reasons for action proper. For by analytically identifying these kinds of reasons with each other, we can explain the normative governance of action by the latter in terms of the normative governance of action by the former. We can summarize this identification by saying that the following is an analytic truth about the relationship between fitting motives and reasons for action:

> **Motivation-Action principle [MAP]**:
>
> Let ϕ be an action or an omission of an action. If it is fitting to be motivated to ϕ, then one has reason to ϕ.[110]

[110] My warrant composition and motivation-action principles closely correspond to Skorupski's partition of his "Feeling/Disposition Principle," (FD): "if there is reason to feel ϕ then there's reason to do that which ϕ disposes to") into (FDF): "If there's reason to feel ϕ there's reason to desire to do that which ϕ characteristically disposes one to desire to do," and (FDD): "If there's reason to desire to do α (or to bring it about that *p*), there's reason to α (to do that which will bring it about that *p*)" (see (Skorupski 1999), especially p. 38, 63, 131, and 174 n24). The main possible differences between my principles and

where its being fitting to be motivated to ϕ includes its being fitting to be intrinsically motivated to ϕ, its being fitting to prefer to ϕ, and its being fitting to intend to ϕ.

3.6. Ethics and Reasons for Action

In light of the relationships we have been establishing between fitting motives and reasons for action, we saw above how fitting attitude analyses of ethical concepts can explain the connection between ethics and rational ends. If facts about what's worth pursuing are identical to facts about what it's fitting to be intrinsically motivated to pursue, and facts about intrinsic ethical status are facts that certain intrinsic motivational states are fitting, then ethical facts are facts about what it's worthwhile to pursue. But we can now say more about the relationship between ethics and reasons for action in light of the connections we have lately seen between reasons for action and reasons for the motivational states that lead us to action.

The first thing to observe is that fitting attitude analyses of ethical concepts, in conjunction with the Warrant Composition Principle and Motivation-Action Principles (and the fact that valenced attitudes involve motivations) give us the result that ethical facts quite generally entail facts about what we have reason to do. Thus, the fitting attitude analysis of a state of affair's goodness that we saw in Chapter 2 entails that if a state of affairs is good, whether intrinsically or simply because it helps lead to something else that is intrinsically good, we have reason to have pro-attitudes towards it. But these pro-attitudes involve motivation to bring the state about when we can. Thus, by the Warrant Composition Principle, if it is fitting to have these pro-attitudes towards a state of affairs, it is fitting for us to be motivated to bring it about when we can. But by the

---

Skorupski's FD principles is my WCP's insistence on ψ's essential involvement as component of ϕ rather than (in his FDF) its mere involvement as what "ϕ characteristically disposes one to desire to do" and my MAP's clarification that the relationship between reasons for motivations and actions is held to hold only between fittingness (and not pragmatic) reasons for motivation and reasons for actions that bring about their objects. But I would think that Skorupski's FD principles might be interpreted along the lines suggested by my warrant composition and motivation-action principles, and that the reasons I give for supposing the latter to be conceptual truths would support such an interpretation of Skorupski's FD principles.

Motivation-Action Principle, if it is fitting to be motivated to bring about the state of affairs when we can, then bringing it about is something we have reason to do when we can. Similarly, the fitting attitude analysis of an act's wrongness we saw in Chapter 2 entails that if an act is morally wrong, whether intrinsically or simply as a way of doing something else that is intrinsically wrong, then it is fitting for us to feel obligated not to perform it. But since this feeling of obligation not to do something involves motivation not to do it, it follows by the Warrant Composition Principle that if it is fitting to feel obligated not to perform the act, it is fitting to be motivated not to perform it. But by the Motivation-Action Principle, if it is fitting to be motivated not to perform the act, we have reason not to perform it.

These entailments reflect the way in which something's ethical status entails the existence of reasons to act in the service of distinctive kinds of ends. We have seen how fitting attitude analyses can understand judgments of intrinsic goodness and wrongness as thoughts about the fittingness of intrinsic desires and feelings of obligation. Now to think something instrumentally good is to think it good on account of its further consequences, or good because it helps bring about something that is intrinsically good. The fitting attitude analyst thus interprets this as a thought about the fittingness of instrumental pro-attitudes, or pro-attitudes towards a state on account of its brining about an outcome that befits intrinsic pro-attitudes. Intrinsic pro-attitudes towards a state involve motivations to bring it about as an end (when one can), and instrumental pro-attitudes towards a state involve motivations to bring it about (again, when one can) as a means of achieving some state towards which one has intrinsic pro-attitudes. So by the Warrant Composition Principle, the fittingness of instrumental pro-attitudes towards a state of affairs entails the fittingness of being motivated to bring it about as a means of achieving an intrinsically good state of affairs. But as we have seen, its being fitting to do something as a means to an end is the same as having reason to do it as a means to that end. So if a state of affairs is instrumentally good, one has reason to bring it about simply as a means of bringing about the intrinsically good state of affairs it is a way of bringing about, and quite independently of its other consequences. By symmetric reasoning, we have it that one has reason to avoid doing what is instrumentally wrong simply as a way of avoiding doing the intrinsically wrong thing that it is a way of doing.

In this way, ethical considerations are genuine reasons to do things simply because they serve distinctively ethical ends. For as we have argued, to be a reason to do something is to be a consideration that counts in favor of preferring and intending to do it, and a consideration does this by helping to make it the case that the action contributes to an end that it is fitting to be motivated to pursue. According to the fitting attitude analyst, a consideration contributes to something's ethical status by showing that thing to befit a certain valenced attitude. Thus, a consideration about a state of affairs, like its providing us with the necessary means to make children happy, contributes to its status as good by helping to make that state a fitting object of pro-attitudes. And a consideration about an action, like its constituting a way of harming innocents, contributes to its status as wrong by helping to make that act a fitting object of prospective guilt-tinged aversion. But valenced attitudes involve motivations to do things, and we have by the Warrant Composition Principle that a consideration which counts in favor of having a valenced attitude equally counts in favor of having the motivations it involves. Considerations that contribute to something's ethical status thus *ipso facto* count in favor of our having the relevant valenced attitudes, having the motivations they involve, and performing the acts that realize the objects of these motivations.

What makes each ethical end distinctive is thus nothing about its content. It is rather something about the kind of attitude the fittingness of which constitutes the rationality of the end. A state of affair's goodness entails that it is fitting to be moved to bring it about out of desire for that state of affairs. An act's wrongness entails that it is fitting to omit doing it out of a feeling of obligation not to do it. Similar remarks could be made for other ethical notions – an act's lowliness entails that it is fitting to be moved to avoid it out of a sense of honor, a person's virtuousness entails that it is fitting to esteem her and be moved to emulate her out of this esteem, an activity's value entails that it is fitting to be moved to engage in it out of appreciation for that activity, and so on. Of course, we can do what we have these reasons to do out of other motives; you can bring about a good outcome because you feel obligated to do so, you can avoid doing what's wrong because you fear getting caught, and so on. Indeed, we have seen that our responding to the reasons we take ourselves to have to do something out of a given fitting attitude does not require us to have that attitude; when our attitudes are recalcitrant we

can respond to our views about these reasons by forming intentions directly. But the idea that it is fitting to do something out of a certain kind of motive stamps the purported reason as a reason of a certain kind, and we can recognize this stamp regardless of the particular considerations that the thinker takes to be reasons of this kind.

An important advantage of individuating different kinds of ethical reasons by the kinds of attitudes out of which it is fitting to act on them is thus that it enables us to make sense of the intuitively wide variety of coherent views about what those reasons are. As we saw in Chapter 2, people can coherently think that all kinds of outcomes are intrinsically good, and people can coherently think that all kinds of acts are morally wrong. Correspondingly, they can coherently think that we have reason to bring about radically different states of affairs on account of their goodness, and that we have reason to avoid radically different acts on account of their wrongness. Any attempt to individuate categories of ethical reasons simply on the basis of content will fail to make sense of these coherent views. For instance, the view that moral reasons are reasons about how to treat others fails to make sense of views about duties to oneself, the view that moral reasons are reasons that are extremely weighty fails to make sense of the view that that reasons associated with honor and aesthetics are at least as weighty as those associated with morality (let alone the common sense view that there are some moral reasons – say not to tell white lies or not to hurt the feelings of the guilty – that are quite weak, even if other moral reasons are not), and so on.

It is important to be able to correctly interpret even those coherent views about moral reasons that are radically false so that we know what they have got things radically wrong about, and what this idea of a moral reason is that we can identify better than they. It is important to understand the questions we are asking when we ask about our moral reasons if we are to make progress on the interesting metanormative questions about how it is that we can identify these reasons. For instance, if someone were to try to identify the very idea of a moral reason with that of a reason to act so as to maximize universal happiness, he would not only beg every important normative question about what our moral reasons actually are. He would also badly misrepresent the deliberative and practical concerns we have when we think about morality. Even if maximizing happiness

is what we have moral reason to do, one certainly does not find that out by finding out what maximizes happiness. One's evidence for such a view stems from basic normative inquiry that seeks to best unify one's non-debunked moral intuitions of all levels of generality. If we want a concept of a moral reason adequate to explain this kind of serious philosophical inquiry, we will need something that can serve as a medium between different possible philosophical positions, and something that can explain the kind of evidence that can be marshaled in favor of a victor.

In the same way, it seems that our moral reasons would still be genuine reasons for action even if they turned out to be reasons to do something other than to maximize universal happiness (or any other particular substantive thing that you like). While we can think that all kinds of considerations are moral reasons for action, taking them to be such is somehow different from thinking that one has other kinds of reasons to do it. One can think that the very same consideration, say that an act will make everyone happier, is a reason to perform it on account of the intrinsic value of each being's welfare and its making the world an intrinsically better place, on account of its being intrinsically wrong not to do so, on account of its making he world intrinsically more beautiful, and so on. These thoughts are all quite distinct. They draw on different sets of intuitions and guide different responses in us. The intuitions to which our theories of what is wrongful answer are those that command our feelings of obligation, and when one acts so as to avoid doing what is wrong, one looks ultimately to these kinds of intuitions to justify one's conduct.

Of course, one might try to deny that moral reasons are (as a matter of conceptual fact, anyway) genuine reasons at all. One might maintain that moral reasons are called reasons by courtesy, and that it is a substantively open question whether they actually count in favor of actions in the same way as, say, our reasons to avoid our own pain or what have you.[111] But for reasons we have seen, this would be inconsistent with the way in which ethical thinking guides deliberation, decision, and action like any other kind of thinking about what to do. Of course, we could use ethical language to express whatever concepts we like, and there might well be occasions when English speakers use terms like

[111] Or one could start playing games with the favoring relation, saying that "of course these considerations count in favor of actions, but what we mean here by 'count in favor' is different from what we mean when we say that the fact that something will avoid your own pain counts in favor of doing it…" and so on.

'moral' to express something that isn't essentially action guiding.[112] But if we were to use the term 'moral' to express something other than the attitude and action guiding notion of what it's fitting to feel obligated not to do, we would fail to use it to express the notion that figures into our serious action-guiding philosophical thinking about morality.

Since what it is fitting to feel obligated not to do bears directly on what to do, this is an important thing to think and talk about in a way other concepts are not. Also, since our basic access to this question of what to do is achieved by the standard sorts of feeling-of-obligation-guiding intuitions that we draw upon in standard philosophical discussions about wrongness, what it is really fitting to feel obligated not to do seems to be the genuine target of this enterprise. Basic questions about what to feel obligated to do, or at least other questions about what valenced attitudes to have, are also evidentially prior to the practical relevance of many related questions, like those which might be suggested by certain features of work by Rawls (1971), Boyd (1988), Brandt (1979), Copp (1995), and Railton (1986). Questions about what makes society most stable, or what everyone would agree to behind a veil of ignorance, or what would make everyone better off (or any other permutation involving society, welfare, and agreement), only seem practically relevant to the extent that it can be independently established that we should care about them, or should feel obligated to abide by certain answers to them, or that we should somehow or other engage our valenced attitudes in how they turn out. So these questions should wait in line with the questions asked by every other substantive ethical theory. Bids for conceptual replacement seem to be bids edge out substantive competitors without argument, and to make us forget the standards with which to evaluate their merits to boot.

It is worth noting, of course, that it is not obvious to people why ethical facts entail facts about reasons for action, and that some people have tried denying that there is any such entailment. But as Falk (1948) observes, we seem to misunderstand the question 'why ought I be moral' if we try to adduce independent, substantive reasons to be moral. For the questioner seems to be interested in understanding why the bare fact

[112] Though most often it probably is parasitic on a use of 'moral' that is action guiding – as when people use 'moral' to mean something like MORAL ACCORDING TO SO AND SO'S VIEWS.

that it is wrong to do something entails that there is reason not to do it. It is consistent with the explanation of why an act's wrongness entails that one has reason not to perform it that it is by no means obvious that the entailment holds. To see it ourselves, we had to find an analysis of moral notions that seemed to best explain their semantic, epistemic, and causal properties, examine how deliberation extends to questions of which ends are worth pursuing and in what this consists, explore how valenced attitudes involve motivational states and how the fittingness of the former entails the fittingness of the latter, and look at how reasons for motives relate to reasons for action. If our account is correct, it is understandable how one of these steps or one part of their combination might not be obvious to someone. So the apparent openness of the question as to why we have reason to be moral is completely consistent with our metaethical account of the answer.

If fitting attitude analyses can in this way be paired with the current understanding of practical reasoning to explain the roles that ethical thought plays in thinking about what to do and guiding action, I submit that any hold-out linguistic intuitions about hand-clasping and such should be jettisoned as so much philosophically uninteresting noise. I take it that these intuitions are already on the ropes – there are far too many intuitions about perfectly coherent ethical thoughts (e.g. that cursing is intrinsically wrong) that look just like them, and there might be a certain kind of failure of imagination causing people to get intuitions of incoherence where they should just have intuitions of wild falsity (e.g. failure to sufficiently imagine if someone *really were* guided in her deliberations and desires by views that played the same roles as assessments of hand-clasping as fitting). But if ethical thought guides our deliberation and action in the form of thoughts about fitting attitudes, what use have we for ethical concepts according to which it is incoherent to attribute intrinsic ethical status to hand-clasping? Why would we have a concept of a good state of the form A STATE BEFITTING DESIRE…UNLESS BY THE WAY YOU THINK THAT STATES OF HAND-CLASPING BEFIT DESIRE. What role would such a restriction play in our deliberations or decisions? It seems that we rule out promoting states of hand-clasping for their own sakes in the same way we rule out promoting states of racial purity for their own sakes: by means of basic normative inquiry that shows these views to be incompatible with a best unification of non-debunked substantive intuitions

about fittingness. I fail to see what could be gained by trying to ridicule substantive views that lose badly by these methods by calling them incoherent, except of course the obscuring of what our own ethical thought is about and what it has to do with our reasons for action.

3.7. Fitting Attitudes and What We Have Most Reason to Do

Let us conclude by taking a closer look at how facts about what it is fitting to be motivated to do determine not only facts about what have *some* reason to do, but facts about what we have *most* reason to do. As we have seen, our judgments about what ends are worth pursuing, in conjunction with our means-ends beliefs, determine a set of actions that are, according to our views, most choiceworthy in the sense that they do best by way of the ends that we judge to be worth pursuing. Correspondingly, what we in fact have most reason to do is what is actually most choiceworthy – what will actually do best by the ends that are actually worth pursuing. Any of these most choiceworthy things is equally something that we have most reason to do or something that is equally rationally permissible.[113] On the other hand, any act that is not among the most choiceworthy acts is rationally impermissible – it is something that we have strictly more reason not to perform; something that we have conclusive reason not to do. It is only in the special case when there is a single most choiceworthy action that we have conclusive reason to perform one act instead of any of its alternatives.[114]

---

[113] I realize that using the language of rational status in talking about what is worth doing, or worth doing in light of all the facts, may be a risky business. Talk of rationality is usually more closely tied to evaluating how well an agent has responded to her reasons given her evidence, her limitations, the burdens she faces, and so on. But here I mean to use talk of 'rationally permissible action' stipulatively to refer an action that is such that one has no more reason not to perform it than one has to perform it. Analogous remarks hold for all the talk in what follows of rational options, mandates, and so on, as well as the talk of rational ends above and throughout.

[114] Though how exactly we carve up acts and their alternatives is somewhat arbitrary and context sensitive. Thus, suppose that we have strictly stronger reasons to keep a promise to stop by Suzy's house some time today than we have not to keep this promise. One way we could describe our situation is one in which we have ever so many permissible options (e.g. going to the movies and then seeing Suzy, seeing Suzy and then going to the movies, running in circles and seeing Suzy at 2:13 PM, …), and only a few impermissible options, namely the ones in which we never stop by Suzy's house. But another way we could describe our situation would be as one in which our only alternatives are (i) stop by Suzy's today, and (ii) not stop by

What I want to examine is what it is for one option to do better than another by way of the ends that are really worth pursuing. Understanding this will enable us to understand what makes an alternative rationally permissible, rationally impermissible, or rationally required. The first thing we need to examine is the distinction among two kinds of ends that are worth pursuing: rationally optional ends that we *may* seek to achieve, and rationally mandatory ends that we *must* seek to achieve. The second thing we will need to examine is the question of the relative importance of different ends. I propose to examine these in turn.

Let us use the term 'goal' to refer to an end at which one aims, or something that one is intrinsically motivated to do or bring about. When it is your goal to do something, omit something, or bring something about, you will be disposed to act in ways that involve or contribute to your doing, omitting, or bringing it about. But of course since you may have many goals, not all of which can be achieved at the same time or with the same plan of action, you can have goals that you are not currently pursuing or intending to achieve. Indeed, you can have some goals that you never actually pursue, because whenever it is possible to pursue them, you pursue other goals instead. Thus, when travelling one of your goals might be to get a good meal, and another might be to get a comfortable night's rest, but if you cannot afford both you might well act with the intention of getting the night's rest, intending to do nothing to realize your goal of getting the meal. Similarly, when you are equidistant between Bugsy and Suzy on your rowboat, you might have saving each of them as a goal, but given that you cannot save both you might form the intention to save only one.

Intuitively, certain goals are such that there are reasons to have them, and you are justified in having them, but it's also O.K. if you don't have them. Goals like these might include doing certain kinds of things full-time or professionally (like philosophy or construction work), playing certain kinds of games or sports (like Go or Basketball), and perhaps listening to certain kinds of music (like Jazz or Classical[115]). One can think, that is, that doing philosophy full time, or playing Go, or listening to Jazz are all things that

---

Suzy's today, where (i) the unique rationally permissible option. But the main thing to keep in mind is that however we carve up the alternatives, there will very often be more than one rationally permissible option.

[115] Let me emphasize that these examples are merely intended to be examples of goals that it seems quite coherent to judge as justified but non-mandatory. I do not wish to pick fights with anyone as to whether listening to Jazz or Classical music is something everyone should want to do.

are well worth wanting and doing as ends in themselves, but that it is also fine to have no desire or inclination to do these things. Let us say that thoughts of this kind are thoughts that the goals in question are *rationally optional*.

Next, there are certain goals that are intuitively such that, at least in certain circumstances, one must have them. Examples like these might include those of taking care of one's health, doing what one can to bring it about that children (and other innocent beings) do not suffer, and keeping certain promises that one has made. One might think that one is not only justified in having goals like these, but moreover that one would be unjustified in failing to have them. Let us refer to thoughts of this kind as thoughts that the goals in question are *rationally mandatory*. The idea here is that it is fitting for you to be motivated to do these things and unfitting for you to fail to be so motivated. Correspondingly, we might say that attitudes which are warranted or fitting in this sense are *rationally mandatory*. Other examples of attitudes that we usually take to be rationally mandatory would include (pace James (1897)) the degree of belief one should have in a proposition given one's evidence.

Thinking an attitude fitting in the sense of rationally mandatory should be contrasted with thinking it fitting in the sense that we think it fitting to be intrinsically motivated to bring about a rationally optional goal. Someone who thinks that playing Go and listening to Jazz are rationally optional ends will think that intrinsic motivations do these things are justified and appropriate but in no way rationally required. Correspondingly, we might say that attitudes that are fitting in this sense of justified but not required are *rationally optional*. Other examples of thinking about rationally optional attitudes would include the way in which we often think about anger towards those who have culpably wronged us. The idea is frequently that one is justified in being angry at people who have culpably wronged you, but that it is also in no way unfitting to fail to have this anger.

There are some important things to clarify about judgments to the effect that an attitude is rationally mandatory. The first concerns the distinction between occurrent and dispositional motivational states. We will often think that a goal is rationally required because we think that a valenced attitude that involves intrinsic motivation to bring it about is rationally mandatory. Thus, we might think that everyone should be averse to

the suffering or innocents or want them not to suffer, and we might think that (unless one are going to keep them anyway) one should feel obligated to keep certain promises. By the reasoning behind the Warrant Composition Principle, these judgments that valenced attitudes are rationally mandatory entail judgments that the intrinsic motivations they involve are also rationally mandatory. Now as we observed in chapter 2, motivational states like desires and emotions involve distinctive subjective experiences, motivations, and patterns of attention direction. But it might well seem unreasonable to demand that people go around constantly with this kind of affect and attentional focus. At the very least people need to sleep. Similarly, the man who must keep a promise a year from now (but might not do so in the absence of feeling obligated) is surely allowed the occasional hour or two of lighthearted entertainment, during which the experience of feeling obligated and thoughts of promise keeping may be far from his mind. So too one does not always have the time to feel the repelling force of aversion to the abhorrent conditions in which impoverished children live; one must at the very least be permitted to focus from time to time on tasks that will enable one to earn one the money that one will use to help them.

We must distinguish, however, between having valenced attitudes in what are standardly called the 'occurrent' and 'dispositional' senses. When one has a valenced attitude in the occurrent sense, one actually experiences the syndromes of phenomenology and directed attention that it involves. Thus, Darwall (1983, 40) gives us a nice example of an occurrent desire attribution in his discussion of a desire to eat pie:

> It may be true of me that were the aroma of fresh apple pie to waft past my nose I would be moved to discover its source and perhaps try to wangle a piece. It does not follow from this, however, that before I smell the pie I desire it or to eat anything at all. Without my encountering its aroma, apple pie, or food in general, might be the farthest thing from my mind or cares (Darwall 1983, 40)

Thus, in the occurrent sense of 'desire' one lacks a desire for pie unless pie is currently on one's mind, one is experiencing a tug in its direction, and so on. Similarly, there is an occurrent sense of 'feeling obligated' in which Jones does not actually feel obligated now to keep his promise unless he is thinking about the promise, experiencing the guilt-tinge of the aversion to failing to keep it, and so on.

But there are equally good senses in which we can attribute valenced attitudes to people who are not currently experiencing the phenomenology or the attentional focus that these attitudes involve when they are had occurrently. Thus, I can truly say that I have been angry at someone for years, and still am, even though I have clearly not been permanently experiencing the phenomenology and attentional focus of anger for years and may well not be experiencing them now. In the same way we can say truly of people who are asleep or unconscious that they desire this or that or feel obligated to do such and so. Thus, if Bill comes from a country where apple pies are rare, but definitely wants to eat some apple pie before he dies, we can truly say of Bill while he is asleep (and not, by the way, dreaming of pies) that one of the things he really wants is to eat apple pie. Or suppose that you are accompanying Smith on a trip to his uncle's for the Winter Holiday. Smith really hates being around his uncle, but he promised he would visit and felt obligated to make the visit. You and Smith have stopped for the night at a hotel, where he is in the room engrossed in watching a horror movie and you are down at the hotel bar telling someone where you are from and where you are going. Even though Smith's thoughts are as far as possible from the promise, the trip, and the guilt-tinge of aversion to failing to make the visit, it seems that you could truly explain to the other person at the bar that Smith feels obligated to visit his uncle.

To judge that a valenced attitude is rationally mandatory is to judge only that one must have it in the dispositional sense. The senses in which everyone should be averse to the suffering of children, and that you should feel obligated to keep certain promises that you might not keep, are those which are compatible with your not currently experiencing the phenomenology and attentional focus of these attitudes. There are, however, important connections between having attitudes in the dispositional and the occurrent senses. To have an attitude in the dispositional sense is, as the name might suggest, to have a disposition to have it occurrently. Now what controls whether this disposition is manifest on any particular occasion has a lot to do with the extent to which one directs one's attention to features that tend to elicit it. Thus, a person with an aversion in the dispositional sense to children's suffering will tend to have the desire occurrently the more his attention is fixed on their suffering, a person with a feeling of obligation to keep a promise in the dispositional sense will tend to feel obligated occurrently the more he

thinks about the promise, and so on. The extent to which one's attention is focused on something is, moreover, itself something that we can assess as warranted or fitting (both in the sense of rationally mandatory and justified but non-mandatory). Thus, we can get to a question about which occurrent attitudes are rationally mandatory by combining that of which dispositional attitudes are mandatory with that of the extent to which one should focus attention on the relevant elicitors.[116]

A second thing that should be clarified is that whether a goal or intrinsic motive is rationally required or mandatory can be a massively circumstance-specific matter. While I often discuss goals that might seem required in most or even all circumstances, this is done only for the sake of simplicity, and should not obscure the fact that a host of considerations can oppose, weaken, or strengthen the case in favor of a given person's having to have any given attitude or goal. Thus, we might think that keeping a promise is usually a rationally required goal, but the fact that the promise has been given under duress might undermine its status as such. In the same way, we might think it rationally mandatory to be averse to causing people to be in pain, but if the pain in question is minor and the person in question has a masochistic desires for it and would enjoy it, it might well be rationally permissible not to be averse to causing the person to experience it. Thus, what is rationally mandatory for one person or for people in one kind of circumstance may well be different from what is rationally mandatory for other people or

[116] I believe that this distinction between questions of what attitudes one should have dispositonally and to what extent one should focus one's attention in ways that cause them to become occurrent is part of the distinction drawn by Gibbard (1990, 126-127) between "how to feel if one's feelings are fully engaged" and "how fully to engage one's feelings in a situation." Gibbard suggests that moral questions are essentially questions about the former. An act's status as blameworthy, for instance, is a matter of its being fitting to feel anger at its author or feel guilt for doing it *if* one's feelings are fully engaged, but the extent to which one's feelings actually ought to be engaged is a question to which an act's status as blameworthy does not speak. As we will see later, there are good reasons to think that certain moral questions do bear on the extent to which one should engage one's feelings, but something is certainly correct about Gibbard's suggestion about the extent to which many ethical questions speak to the issue of "how to feel if one's feelings are fully engaged." That an act is blameworthy in many ways leaves open the question of the extent to which any given person should engage her attention in ways that give rise to occurrent anger.

I think, however, that silence on questions of when one's dispositional attitudes must become occurrent is not the whole of the story about the extent to which ethical questions can fail to settle how one must feel in one's circumstances, and I suspect that Gibbard was getting at more than this with his distinction between "how to feel if one's feelings are fully engaged" and "how much to engage one's feelings." For it seems plausible (and at least coherent) to think that certain acts are blameworthy but that one is quite permitted to fail to have even dispositional feelings of anger at their authors. Blameworthiness thus seems to speak to the status of dispositional anger as justified rather than rationally mandatory.

those in other circumstances. Particularists need in no way oppose the idea that certain goals and attitudes are rationally required or mandatory in the relevant circumstances.

Now if a goal is rationally optional, the fact that an action will contribute to its achievement can count in favor of the action in the sense of contributing to its status as justified or permissible. Thus, it might be rationally impermissible to cause oneself pain for absolutely no good reason, but the fact that one needs to do certain painful things in order to pursue a rationally permissible goal (like becoming a great athlete) can make it permissible to do them. But since by definition nothing favors one's having a rationally optional goal over one's not having it, the fact that an action will contribute to such a goal cannot by itself count against failing to do it in the sense of contributing to its status as rationally impermissible. If listening to some jazz is a merely rationally optional goal, then the fact that one must go to a given café to listen to some jazz cannot make it impermissible not to go to the café, since it is perfectly permissible not to be seeking out jazz in the first place.

On the other hand, if a goal is rationally mandatory, the fact that an action will contribute to its achievement contributes both to the rational permissibility of the action and to the rational impermissibility of not performing the action. Thus, it might be rationally impermissible to painfully stick pieces of metal into your skin for no good reason. If having a body piercing is a rationally optional goal, then the fact that you need to painfully stick a piercing into your skin in order to have a body piercing can contribute to the rational permissibility of doing so, but it cannot by itself contribute to the rational impermissibility of not sticking the body piercing into your skin. But if preserving your health is a rationally *mandatory* goal, then the fact that you need to stick a needle into your skin in order to get a shot that will keep you healthy not only contributes to the permissibility of sticking yourself with the needle, but in fact contributes to the rational impermissibility of failing to stick yourself with the needle. In the same way, it might be rationally impermissible to ruin your best clothes for no good reason. If making an avant-garde art project out of your best clothes is a rationally optional end, then the fact that you need to cut them up for the project can contribute to the permissibility of cutting them up, but it cannot on its own contribute to the impermissibility of failing to cut up your clothes. If, on the other hand, bringing about the survival of a child is a rationally

mandatory goal, then fact that you need to wade into a muddy pond to save her from drowning not only contributes to the permissibility of ruining your clothes by wading in, but in fact contributes to the rational impermissibility of failing to do so.

Since a goal's status as optional or mandatory consists in its being rationally optional or mandatory to be intrinsically motivated to achieve it, we have then the following relationships between fitting motives and reasons for action. If intrinsic motivation to do *X* is fitting in the sense of rationally optional (justified but not mandatory), then the fact that doing *Y* helps one to do *X* can contribute to one's having most reason to do *Y* (or no stronger reason not to do *Y* than to do *Y*), but it cannot by itself contribute to one's having conclusive reason to do *Y*. If, on the other hand, intrinsic motivation to do *X* is fitting in the sense of rationally mandatory, then the fact that doing *Y* helps one to do *X* can contribute to both one's having most reason to do *Y* and to one's moreover having conclusive reason to do *Y*, or strictly more reason to do *Y* than not.

An important feature of motivations to do things is that they come in degrees: one can be more or less motivated to do something, and one can be more strongly motivated to do one thing than one is motivated to do another. Talk of preferring one outcome to another is a special case of talk about what one is more strongly motivated to do. When one prefers outcome S to outcome S', one is more strongly motivated to bring about S than one is to bring about S'. When this is so, one will tend to do what (according to one's beliefs) will bring bout S rather than what will bring about S' when one is faced with a choice between the two. Now as we have seen, preferences over states of affairs are only one kind of motivational state; we can also have motivations to do or omit certain things, which cannot be reduced to motivations to bring about certain outcomes or states of affairs. So more generally, when one is more strongly motivated to do *X* than one is to do *Y*, one will tend to do what (according to one's beliefs) will contribute to one's doing *X* rather than what will contribute to one's doing *Y* when one is faced with a choice between them.

We can think of a motivation to do *X* as coming with a certain strength, which can weigh against motivations to do things that one takes to be incompatible with doing *X*. The strength of one's motivation to do *X* can also combine with the strength of other

motivations to do *X* to outweigh motivations to do things incompatible with *X* that might have defeated the initial motivation to do *X* on its own. Thus, Jones the mobster might be somewhat motivated not to kill Smith, another mobster, out of fear of getting caught, which can weigh against his motivation to kill him out of anger. But if the chances of getting caught are very small and Jones see that by killing Smith he can marry his spouse, his motivation not to kill him out of fear of getting caught might be too weak to resist these motivations on its own. But in alliance with his feeling of obligation or prospective guilt-tinged aversion to killing Smith, Jones's motivation not to kill Smith out of fear of getting caught might be strong enough to overpower his motivations to kill him out of anger and so as to marry his spouse. And so on.

Intrinsic motivations in particular will be had with degrees of strength. An agent's intrinsic motivation to do *X* will be stronger than her intrinsic motivation to do *Y* when the former outweighs the latter, giving rise to a tendency for her to do what, according to her beliefs, will bring about *X* rather than *Y* when she takes herself to be in a situation in which she cannot bring about both. So in addition to questions about the fittingness of an intrinsic motivation to do *X*, understood as questions about whether it is fitting to have an intrinsic motivation to do *X* of some strength or other, we can ask questions about whether it is fitting to be intrinsically motivated to do *X* with a particular degree of strength, and whether it is fitting to be intrinsically motivated to do *X* more strongly than one is intrinsically motivated to do *Y*.

I have argued that identifying judgments about what ends are worth pursuing with judgments about what ends it is fitting to be intrinsically motivated to achieve can explain the way in which the former are shaped by basic normative inquiry and the way in which they normatively guide our motivations and actions. Implicit in this was actually an argument for identifying judgments about the extent to which an end is worth pursuing with judgments about the fittingness of particular degrees of strength of intrinsic motivation to achieve it. For in our case of the agent convinced to be vegan by philosophical arguments, she might well have had *some* intrinsic aversion to causing suffering and death to non-human animals. It was simply that before she did her basic normative inquiry, this aversion was weaker than her desire to eat certain things. What the argument convinced her of was that it is fitting to have an aversion to contributing to

the harm of non-human animals that is stronger than any desire she might have to do things that harm them (much in the same way she has – and thinks she should have – an aversion to contributing to the harm of mentally comparable humans that is stronger than her desires to eat the relevant foods). Similarly, what the formerly racist agent was convinced of was that she should be *just as averse* to harming members of other races as she is to harming members of her own. Reflective equilibrium methods thus seem to deliver verdicts on the strengths or relative strengths of the intrinsic motives it's fitting to have, not just what intrinsic motives it's fitting to have to some degree. And these judgments about the appropriate strengths of intrinsic motives have a propensity to directly generate intrinsic motives of the corresponding strengths. The agent convinced to be vegan did develop an aversion to harming non-human animals that was stronger than her desire to eat certain foods, and the racist did end up having a stronger aversion to harming members of other races (which, in the absence of all recalcitrance, would be as strong as his aversion to harming members of his own).

In effect, then, we have argued that to judge that the end of doing *X* is more important than end of doing *Y* is to judge that it is fitting to have a stronger intrinsic motivation to do *X* than to do *Y*. Now as we have lately seen, there are different senses in which it can be fitting to be intrinsically motivated to do something, corresponding to different ways in which doing it is an end worth pursuing. An intrinsic motive can be fitting in the sense of rationally optional, in which case the realization of its object is a rationally optional goal, or the intrinsic motive can be rationally mandatory, in which case the goal of realizing its object is rationally mandatory. This distinction between rationally optional and rationally mandatory attitudes also applies to intrinsic motives of a given strength and consequently the relative strengths of our intrinsic motives. Thus, it might be fitting in the sense of rationally optional to have a stronger intrinsic motivation to play Go than one has to play basketball, and also fitting in the sense of rationally optional to have a stronger intrinsic motivation to play basketball than to play Go. But it might be rationally mandatory to have a stronger intrinsic motivation to preserve one's health than one has to avoid the pain of a needle-prick - fitting to have intrinsic motivations with these strengths, and unfitting not to. Similarly, it might be fitting in the

sense of rationally mandatory to have a stronger intrinsic motivation to save a child from drowning than one has to avoid dirtying one's best clothes.

As we saw with the question of whether a goal is rationally optional or rationally mandatory, whether it is rationally optional or rationally mandatory to pursue one goal at the expense of another bears on whether one is rationally permitted or moreover rationally required to do so. Thus, in a given situation, it might be rationally optional to have a stronger intrinsic motivation to be a bit healthier than one might have to read a bit more about something interesting. This could make it all things considered rationally permissible to spend more time running and less time reading. But given that it is not rationally mandatory to have this greater concern for one's health, the conflict between additional health and additional reading will not automatically make it rationally impermissible to spend more time reading and less time running. On the other hand, if it is rationally mandatory to have a stronger intrinsic motive to have the health benefits of a vaccination than one has to avoid getting stuck with the vaccinating needle, then the conflict between the avoiding the prick and having the vaccination may make it all things considered rationally impermissible forego having the vaccination.

Of course, to fully answer the question of whether a given action is, all things considered, rationally permissible or rationally impermissible, we to ask how it affects the realization of every goal that is worth pursuing. If there are more than two goals at stake in a situation, we need to ask about the fittingness of having a set of intrinsic motives with certain combined strengths that outweigh certain other combined strengths. Thus, by moving to a new town one might be able to have a more interesting job and get one's child into a somewhat better school, but one might not be able provide one's aging parent with the same standard of care. If one is justified in having intrinsic motivations to have the more interesting job and to have the somewhat better schooling for one's child that have a combined strength greater than one's intrinsic motivation to provide one's mother the somewhat better standard of care, it will be rationally permissible to make the move. If it is simply rationally optional to have motivations of these strengths, will be rationally permissible to move but not rationally impermissible to stay put. But if it is rationally mandatory to care more about the job and the somewhat better education

put together than it is to care about the somewhat better nursing on its own, then it will be rationally impermissible not to make the move.

In general, we might try saying the following for any action *X*. Let $S = \{g_1, g_2, .., g_n\}$ be the set of goals satisfied by one's doing *X*, and let $N = \{g_1', g_2', .., g_k'\}$ be the set of goals that are not satisfied by one's doing *X*, but that would be satisfied by one's not doing *X*. Let $M(S)$ be a set consisting of intrinsic motivations to achieve each of the elements of *S*, such that $M(S) = \{m(g_1), m(g_2), \ldots, m(g_n)\}$, where $m(g_i)$ is an intrinsic motivation to achieve g_i. Similarly, let $M(N) = \{m(g'_1), m(g'_2), \ldots, m(g'_k)\}$be a set consisting of intrinsic motivations to achieve each of the elements of *N*. Then:

(1) one has *most reason* to do *X* (i.e. it is rationally permissible to do *X*) iff one is *justified* in having sets of intrinsic motivations $M^*(S)$ and $M^*(N)$ such that the combined strengths of $M^*(S)$ are greater than or equal to the combined strengths of $M^*(N)$,

(2) one has *most* reason to do *X* but *not* conclusive reason to do *X* (i.e. it is rationally permissible both to do *X* and not to do *X*) iff it is *rationally optional* for one to have sets of intrinsic motivations $M^*(S)$ and $M^*(N)$ such that the combined strengths of $M^*(S)$ are greater than or equal to the combined strengths of $M^*(N)$ [*or* if it is rationally mandatory to for one to have $M^*(S)$ and $M^*(N)$ such that the combined strengths of $M^*(S)$ are exactly equal to the combined strengths of $M^*(N)$], and

(3) one has *conclusive reason* to do *X* (i.e. it is rationally impermissible to fail to do *X*) iff it is *rationally mandatory* for one to have sets of intrinsic motivations $M^*(S)$ and $M^*(N)$ such that the combined strengths of $M^*(S)$ are greater than the combined strengths of $M^*(N)$.[117]

[117] Now in many cases, whether an action achieves a goal depends to a large extent on what else one does. Doing one thing may definitely frustrate one goal, but achieve another goal only if one does something else later. For most of my purposes here I can assume that the picture is static – that all relevant goals achieved by an act are definitely achieved by it or not. But it is worth looking at how we could handle the general case.

To ask about whether it is permissible to perform an act that may or may not achieve certain goals depending upon what else one does, we need to ask about how it could be combined with other acts in a bigger course of action. How precisely an act can be combined with other acts is typically a matter of fact beyond our ken. But we do have a rough and reliable way of estimating this, and for our present purposes it will do to stick with questions of the objective reasons that we are trying to estimate. In this spirit, let us

This notion of an act's achieving the goals that it is fitting to be most strongly motivated to achieve gives us an answer to our question about what it is for an act to "do best by" the ends that are worth pursuing. An act is objectively choiceworthy, or such that it is fitting to prefer to perform it,[118] to the extent that it actually realizes the goals that it is fitting to be most strongly motivated to achieve. The most choiceworthy acts are those which achieve what it is fitting to be intrinsically motivated to achieve on balance. That is to say, they achieve sets of goals which are such that it is fitting to have intrinsic motivations to achieve them with combined strengths that are at least as great as the combined strengths of one's intrinsic motivations to achieve any alternative set of goals. Since the most choiceworthy acts are the set of all those which are rationally permissible, it is rationally mandatory to prefer them to all other acts. For any two most choiceworthy acts, it will be either rationally optional to prefer one to the other or it will be rationally mandatory to be completely indifferent between them. The former will be the case if each serves a set of goals that it is rationally optional to be intrinsically motivated on balance to achieve in preference to those served by the other, while the latter will be the

---

consider such things as "life-plans:" fully determinate courses of action that one could take from the time of present decision to the end of one's life. The rational permissibility of doing something now will thus be a matter of how well the life plans that are consistent with one's doing it contribute to every goal worth pursuing.

More formally, we might try saying the following for any action X. Let $C = \{P_1, P_2 \ldots, P_n\}$ be the set of life-plans consistent with one's doing X, and let $I = \{P_1', P_2', \ldots, P_m'\}$ be the set of life-plans inconsistent with one's doing X. In comparing a given life-plan consistent with X, P_i, with a given life-plan inconsistent with X, P_j', let the satisfaction set of P_i relative to P_j, $S_{ij} = \{g_1, g_2, .., g_r\}$, be the set of goals satisfied by P_i that are not satisfied by P_j', and let $N_{ij} = \{g_1', g_2', .., g_s'\}$ be the set of goals satisfied by P_j' that are not satisfied by P_i'. Let $M(S_{ij})$ be the set consisting of intrinsic motivations to achieve each of the elements of S_{ij}, and let $M(N_{ij})$ be the set consisting of intrinsic motivations to achieve each of the elements of N_{ij}. Then:

(1') one has *most reason* to do X iff there is some element of C, P_a, such that for every element of I, P'_k, one is *justified* in having intrinsic motivations $M^*(S_{ak})$ and $M^*(N_{ak})$ such that the combined strengths of $M^*(S_{ak})$ are greater than or equal to the combined strengths of $M^*(N_{ak})$,

(2') one has *most* reason to do X but *not* conclusive reason to do X iff there is some element of C, P_a, such that for every element of I, P'_k, one is justified in having intrinsic motivations $M^*(S_{ak})$ and $M^*(N_{ak})$ such that the combined strengths of $M^*(S_{ak})$ are greater than or equal to the combined strengths of $M^*(N_{ak})$, but there is no element of C, P_b, such that for every element of I, P'_k, it is rationally mandatory to have intrinsic motivations $M^*(S_{bk})$ and $M^*(N_{bk})$ such that the combined strengths of $M^*(S_{bk})$ are greater than the combined strengths of $M^*(N_{bk})$

(3') one has *conclusive reason* to do X iff there is some element of C, P_b, such that for every element of I, P'_k, it is rationally mandatory to have intrinsic motivations $M^*(S_{bk})$ and $M^*(N_{bk})$ such that the combined strengths of $M^*(S_{bk})$ are greater than the combined strengths of $M^*(N_{bk})$.

[118] where 'objectively choiceworthy or fitting' is again choiceworthy or fitting in light of the facts of the case rather than our best evidence

case if it is rationally mandatory to be just as strongly intrinsically motivated to achieve the goals served by the first as it is to be motivated to achieve the goals served by the second. There will be a unique most choiceworthy act if and only if it serves goals which it is rationally mandatory to be intrinsically motivated to achieve on balance in preference to those achieved by all alternatives. If there is more than one most choiceworhty act, it will be rationally optional to intend to perform either, and if there is a unique choiceworthy act it will be rationally mandatory to intend to perform it.

So whether we are talking about intrinsic motivations, preferences over acts, or intentions, we have a relationship between what it is fitting to be most strongly motivated to do and what we have most reason to do. To be intrinsically motivated on balance to pursue one set of goals in preference to another is to be most strongly intrinsically motivated to do whatever will, according to one's beliefs, bring about the first in preference to the second. If it is fitting to be most strongly intrinsically motivated to bring about a set of goals, it is fitting to prefer, or be most strongly motivated to perform, whatever action will about this set of goals, and as a result to be most strongly motivated to perform it by an intention to perform it. These relationships hold for both fittingness in the sense of rational optionality and rational mandatoriness, as summarized in the following principle that is, if the foregoing is correct, an analytic truth:

Most-Motivation-Action Principle:

(1) If one is justified in being most strongly motivated to do *X*, then one has most reason to do *X* (i.e. it is rationally permissible for one to do *X*),

(2) If it is rationally optional for one to be most strongly motivated to do *X*, or rationally mandatory for one to be indifferent between doing *X* and not doing it, then one has most reason to do *X* but not conclusive reason to do it (i.e. it is rationally permissible both to do *X* and not to do *X*), and

(3) If it is rationally mandatory for one to be most strongly motivated to do *X*, then one has conclusive reason to do *X* (i.e. it is rationally impermissible for one to fail to do *X*).

Chapter 4

The Demands and Recommendations of Morality

We saw in Chapter 2 how analyses of ethical concepts, including moral concepts like BLAMEWORTHINESS and WRONGNESS, in terms of the fittingness of valenced attitudes like guilt, anger, and feelings of obligation, can explain their central epistemic, semantic, and attitude-regulatory features. We also saw in Chapter 3 how these analyses can be combined with an adequate understanding of the relationship between reasoning about which attitudes are fitting and reasoning about what to do in order to explain how moral judgments guide practical reasoning and action, and what morality has to do with one has reason to do. What I would like to do in this chapter is to begin to develop a fuller understanding of our moral concepts, how they relate to our reasons for action, and how they relate to each other.

I will first explore a set of moral concepts that we have not yet discussed in so much detail, namely those of moral goodness, or estimability, and moral badness, or disestimability. We will see how the status of an action as morally estimable guarantees that we have a distinctive kind of moral reason to perform acts of its kind, but that – almost paradoxically – performing these acts for these reasons tends to disable our action from itself counting as morally estimable. This forces us to pay attention to the distinction between an attitude's fittingness and an attitude's estimability. Also, we often think that morally good action is motivated by feelings of obligation, which often enough are not only estimable but in fact fitting or justified. This forces us to attend to the distinction we discussed in Chapter 3 between attitudes that are justified but rationally optional and attitudes which are rationally mandatory. The distinctions between estimable, optional, and mandatory attitudes are, as we shall see, important for evaluating

certain criticisms that might be made of fitting attitude analyses of moral wrongness and moral blameworthiness.

Next, I will take a closer look at our concept of moral wrongness, and take up the issue of how an act's wrongness relates to our having conclusive reason not to perform it. As we shall see, to think an act morally wrongful is to think that, unless one is going to refrain from performing it anyway, it is rationally mandatory to feel obligated not to perform it. We shall also observe a non-standard feature of feelings of obligation – namely that they can only be rationally mandatory if we are unjustified in being more strongly motivated to act to the contrary. Thus, whether there is a rational mandate for feeling obligated to do something is sensitive to whether the *pro-tanto* reasons to feel obligated to do it (like perhaps *I've promised to do it* or *She'll die if I don't do it*) are strong enough to outweigh other considerations in determining what one may be motivated to do on balance. Since feelings of obligation to do something can only be mandatory if it is mandatory to be most strongly motivated to do it, we will have by the Most-Motivation-Action Principle that a rational mandate for feeling obligated to do something entails that one has conclusive reason to do it. So, since an act's status as morally wrong is a matter of the existence of a rational mandate for feeling obligated not to perform the act, an act's moral wrongness entails the existence of conclusive reasons not to perform it.[119]

Having discussed what is morally good and what is morally required, I will turn to the topic of supererogatory action, or action above and beyond the call of moral requirement. I will explore how the notion of supererogatory action seems to differ from both morally good action and various things we see in our discussion of morally required action. I will argue that central to the notion of supererogatory action is that of its being

[119] Actually, as we shall see, there is a mandate to feel obligated not to do wrong only if one is not going to omit the wrongful action anyway. To see how in general the wrongness of an act entails the existence of conclusive reason not to perform it, we will need to consider the case in which one is going to perform the act anyway. One can do so out of either a justified or an unjustified motive. If one's motive is justified, well enough, if one's motive is unjustified, one has most reason not to have it, and thus to be such that the rational mandate to feel obligated not to do it applies (in which case one has conclusive reason to be most strongly motivated not to do it for the reasons mentioned in the text). Thus, if an act is morally wrong, one is only justified in being in some state of overall motivation not to do it (which may or may not involve feelings of obligation not to do it), which is to say that it is rationally mandatory to be most strongly motivated not to do it. It follows from this, by the Most-Motivation-Action Principle, that an act's wrongness entails the existence of conclusive reason not to perform it.

fitting but not mandatory to feel obligated to do something. Because supererogatory acts befit, but do not mandate, feelings of obligation to perform them, an act's status as supererogatory entails that we have some reason to perform it, but it does not entail that such reason is conclusive.

Finally, I will return to our concept of moral blameworthiness and the fittingness of guilt, outrage, and resentment. I show such an analysis enables us to make sense of retributivist thinking and perhaps some more general thoughts about discounting the interests of the guilty. As species of anger, outrage and resentment involve motivations to behave punitively towards their objects. The fittingness of these attitudes, in conjunction with the Warrant Composition and Motivation-Action Principles, thus entails that we have reason to behave punitively towards their objects. At the same time, since guilt involves motivations to make amends and accept punishments, its fittingness entails the existence of reasons to do so. In conjunction with the foregoing account of moral wrongness, this can help explain why, if someone really is morally blameworthy, we might be morally permitted to do things to her (and she would be morally required to accept that we do them) that we would not be allowed to do to the innocent, like to imprison her when doing so will serve some deterrent or reforming purpose. For our general moral reasons not to harm someone or restrict her autonomy might overwhelm our reasons to do so in the service of a greater good and enforce a rational mandate to feel obligated to do so. But if the person has culpably offended, our justified anger might make it the case that we are no longer required to be most strongly motivated not to harm her, removing the rational mandate for feeling obligated not to do so.

There are, however, important questions about whether outrage and resentment really could be justified in these ways. Reflection on the fact that the behavior (or objective chances of behavior) of agents are determined by prior events, laws, and objective chances that are beyond their control might convince us that it is unfair to be moved to harm them on account of what they do. This would entail that no agent around here (and perhaps no metaphysically or even conceptually possible agent) can actually be the fitting object of anger, and that none of us can actually do wrong in a culpable or blameworthy way. Settling whether this is actually so is squarely beyond the scope of this dissertation. But what is important to note is just which parts of our moral thinking

this would and would not affect. If I am correct about our moral concepts and their connection to reasons for action, then universal exculpation on account of the non-existence of libertarian free will would affect the applicability of retributivist thinking to the actual world. If no one is really culpable, we cannot discount anyone's interests simply on account of the justifiability of anger and its motivations to punish. But the fact that anger (and guilt) are never really fitting should in no way affect our confidence that certain things really are morally wrong, estimable, disestimable, and supererogatory. For these concepts require only that feelings of obligation, moral esteem, and moral disesteem be fitting, and their fittingness in no way requires the fittingness of guilt or anger at anyone. Universal exculpation would thus leave completely intact the reasons we have not to do wrong, to do what is good, and to avoid doing what is bad. The absence of libertarian free will might thus threaten our ability to hold each other morally accountable, but it would in no way threaten our reasons to be moral.

4.1. MORAL GOODNESS and MORAL BADNESS

Perhaps the first thing to notice about our notions of MORAL GOODNESS, or MORAL ESTIMABLILITY, and MORAL BADNESS, or MORAL DISESTIMABILITY, is the variety of kinds of things that can fall under them. As I suggested above, we can assess actions, motivations, character traits, or agents as morally good or bad. Moreover, when we do assess actions as morally good or bad, we need to know in addition to other features of an agent's circumstances the motivational state out of which she performs the action. Thus, giving a large percentage of one's income to the poor out of a feeling of care for the poor or out of a feeling of obligation to alleviate their suffering are prime candidates for morally good conduct. But performing the exact same act simply out of a motive to further ingratiate oneself with a potential romantic partner, or a motive to further enrich oneself in the belief that big giving will result in even bigger returns from gratitude and reciprocation later down the line, are not. Similarly, failing to sacrifice one's life to save another innocent person merely out of an aversion to make the supreme sacrifice is presumably not morally bad, but failing to sacrifice one's life to her merely out of a

sadistic desire to see her suffer and die (where we may assume one is indifferent to dying oneself – say due to a bout of clinical depression) is a prime candidate for being so.

This is in sharp contrast to how things stand with our notions of MORAL WRONGNESS, and indeed even MORAL BLAMEWORTHINESS. Only actions and omissions of actions can be assessed as morally wrong, and only such actions or omissions and those who perform them can be assessed as blameworthy.[120] Moreover, although the (subjective) wrongness of an action or omission is sensitive to many factors of the actor's circumstances such as her information, and the blameworthiness of an action or omission is sensitive to yet further features of the actor's circumstances such as her emotional state, these assessments do not essentially depend upon the particular motivations out of which the action was performed. So for instance, if one refrains from harming others merely out of a desire not to get caught, one will refrain from doing something morally wrong and blameworthy every bit as much as someone who so refrains out of feelings of care for the potential victim or feelings of obligation not to harm her. Similarly, in the above cases of those who give to the poor out of good and non-good motives, and those who refrain from sacrificing their lives for others out of non-bad and bad motives, the actors do not differ as to whether they do moral wrong or something morally blameworthy. If both those who give to the poor out of care or feelings of obligation and those who give to the poor out of a desire for ingratiation with others or gain are in the same informational circumstances, circumstances of wealth, etc. they will be morally obligated to give the exact same amount, and either both will meet that moral obligation (by giving a sufficient percentage) or neither will. Similarly, if they are alike in such matters as levels of emotional agitation, temptation to do otherwise, etc. they will both be such that their failures to give the sufficient amount is (or would be) blameworthy - or

[120] We can of course hold to be blameworthy a person's role in getter herself to have or refrain from having a particular character trait or motive (e.g. by cultivating or failing to cultivate it). There may be a way of speaking about such cases according to which we do say things like "he's to blame for feeling or being that way" (which might not be unlike a common mode of speech in which we call pragmatic reasons to get ourselves to have certain attitudes "reasons to have the attitudes"). But however we choose to speak, the object of our assessment in such cases is still the person's acts or omissions on the way to the result of her having or failing to have the attitude or trait rather than her having or failing to have the attitude or trait *per se*.

neither will. And identical remarks hold for the cases of those who fail to sacrifice their lives out of desires not to lose everything and sadistic desires.[121]

Any analysis of MORAL GOODNESS and MORAL BADNESS should thus accommodate and if possible explain these features of the concepts and how they are different from our notions of MORAL WRONGNESS and BLAMEWORTHINESS. Here is a way in which I think this can be done. Begin by noting that, for the concepts in which we are interested, synonyms for 'moral goodness' and 'moral badness' include 'morally estimable' (as well, perhaps, as 'morally admirable') and 'morally disestimable'. Next, recall the suggestion of Brandt (1946) that ""*X* is *Y*-able" means that "*X* is a fitting object of *Y*-attitude (or emotion)"." It seems that there are such attitudes as esteem and disesteem, and moreover that they can take as objects exactly the variety of things that it looks like we can assess as morally good and bad – actions, motives, character traits, and agents. Moreover, just as we assess actions as motivated in a particular way as morally good or bad, so too we seem to esteem and disesteem actions as motivated in particular ways. It seems does not seem that we can esteem the act of giving a large percentage of one's income or disesteem failing to sacrifice one's life for another person in abstraction from how these acts were motivated – we appear to have to draw on representations of whether these acts were motivated by something like feelings of obligation, desire for gain, sadism, etc. in order to admire or look down on their performance. This suggests that the following fitting attitude analyses of MORAL GOODNESS and MORAL BADNESS would help explain the kinds of things that can fall under these concepts:

[121] One might think that some further elaboration would be necessary to make these cases comparable as to what is morally demanded, since it might seem that the depressive may have less to lose by dying and thus will be more likely to be morally required to sacrifice his life. While it would of course be possible to fill in the details to remove any such appearance however one likes, I do not think that this will in general be necessary, since often depressives have just as much to lose as others (and only an insidious conflation of bare current mood with welfare or future directed intention with psychological connection to future would suggest otherwise). What might be true, however, is that a depressive - merely on account of being such - may be less likely to be tempted not to sacrifice his life when he is morally required to do so, and thus more likely (at least on *this* account) to be blameworthy for failure to do so should such a thing be morally required.

Fitting Attitude Analysis of Moral Goodness / Estimability:

Let Φ be an action motivated in a particular way, a motive or motivational state (like a particular desire or emotion), a character trait, or an agent. To judge that Φ is morally good or estimable is to judge that it is fitting to morally esteem Φ.

Fitting Attitude Analysis of Moral Badness / Disestimability:

Let Φ be an action motivated in a particular way, a motive, a character trait, or an agent. To judge that Φ is morally bad or disestimable is to judge that it is fitting to morally disesteem Φ.

These analyses draw on particular kinds of attitudes, 'moral esteem' and 'moral disesteem'. In Chapter 2 we argued in general against judgmentalism about valenced attitudes, but one might be skeptical about the possibility of distinguishing moral esteem and disesteem from other kinds of esteem without invoking the concepts of moral goodness and badness that we are here trying to analyze. It seems, however, that these attitudes involve distinctive phenomenal, motivational, and attentional elements that are quite independent of judgments about or representations of moral goodness and badness.

To morally esteem something involves having a feeling towards it with a phenomenal character that we might try to describe in such terms as "looking up to it," "feeling impressed by it," "standing in awe of it," etc. This phenomenology may be manifest in a particular kind of attentional focus that is characteristic of wishfully imagining or fantasizing about doing or being like the object of the esteem. Thus in morally esteeming something, one may tend to fantasize about or wishfully imagine performing the esteemed action, being in possession of the esteemed motive or trait, or being alike in these ways to the esteemed agent (cf. Velleman 2002). Moreover, the kind of esteem that seems related to assessments of moral goodness seems to feel different from other kinds of admiration. It feels something like an impartial version of gratitude, or has a kind of moralistic tincture associated with feeling like thanking the person whose act, motive, trait, or self is the object of the esteem for doing what she has done or for being as she is. In this way this kind of esteem may be contrasted with those associated with accomplishments that have little to do with morality, such as proving a difficult

theorem, being particularly disciplined in dieting, showing courage in a pride fight, and so on.

Along with these phenomenal features, morally esteeming an agent or her character, motive, or action itself involves motivational elements as well, most notably a motivation to emulate the object of the esteem. Emulation here should be understood in purely behavioral terms.[122] That is, when one esteems an agent's performing of a particular action, one is motivated to perform the action oneself, and when one esteems an agent, character trait, or motive, one is motivated to behave like the agent or one with the trait or motive that is the object of one's esteem.

Similarly, to morally disesteem something involves having a feeling towards it with a phenomenal character that we might try to characterize as "looking down upon it," "feeling shocked by it," "feeling revolted by it," etc. This may be manifest in one's tending to wishfully imagine or fantasize about doing or being *un*like the object of the disesteem. One may thus tend to fantasize or wishfully imagine oneself refraining from performing the disesteemed action in similar circumstances, as being without the disesteemed motive or trait, or as being unlike in these ways to the disesteemed agent.

Just as moral esteem seems to feel different from other kinds of esteem, the kind of disesteem related to assessments of moral badness seems to feel different from other kinds of disesteem. Thinking about the peculiar phenomenology of moral disesteem may be complicated by the fact that there are other negatively valenced moral emotions like guilt, outrage, and resentment, which are associated not with MORAL BADNESS but with MORAL BLAMEWORTHINESS.[123] But just as we have attitudes of gratitude which are associated with morally good conduct that affects us personally, so too we might seem to have a distinctive phenomenology that tends to be associated with conduct that is morally bad – but not necessarily wrongful or blameworthy – that affects us personally. Perhaps we would characterize this as "feeling screwed over," "feeling like we have been given a

[122] My use of 'emulation' is thus different from that of Velleman (2002), who uses 'emulation' to mean not just behaving alike to the object of the emulation but also having attitudes (e.g. emotions) and motivations (i.e. where these are individuated not just in terms of what they are motivations to do but also what kind of motivational state – e.g. emotion or desire in the ordinary English sense – they are a part of).

[123] Still, I think that (among other things) my above remarks about the distinctness of the things that can fall under these concepts (including how the falling of an action under the former but not the latter concept depends upon the particular motive out of which the action was performed *per se*) should be sufficient to convince us that these are indeed distinct concepts.

raw deal," "feeling slighted," "feeling harshly dealt with," etc. Thus, we might try characterizing moral disesteem as akin to an impartial version of feeling "screwed over," "feeling like one has been given a raw deal," "feeling slighted," "feeling harshly dealt with," etc. In this way it has the kind of moralistic tincture associated with feeling as though the person whose act, motive, trait, or self is the object of the disesteem is a "jerk" for doing what he is did or being as he is, or that the disesteemed act was a "shabby" thing to do. In these ways moral disesteem feels distinct from the kinds of disesteem associated with personal failings that have little to do with morality, such as failing to get simple sums right due to being too lazy to expend the mental energy, indulgently overeating, backing down in a personal fight merely out of cowardice, and so on.

Just like morally esteeming something, morally disesteeming something involves having certain motivations. In particular, it involves motivation to *dis*emulate, or to behave in ways *un*like, its object. Thus, when one disesteems an action one is motivated to avoid performing similar actions oneself. When one disesteems an agent, character trait, or motive, one is motivated to behave unlike the agent or one with the trait or motive.

Observing these features of moral esteem and disesteem helps us to see how they are distinct from judgments about moral goodness and badness. For we can make judgments of moral goodness and badness without these attitudes, and we can have these attitudes recalcitrantly, or in conflict with our judgments of goodness and badness. One can judge that something is morally good or bad without the affect, motivation, or attention direction involved in morally esteeming or disesteeming it. In a state of depression or mental exhaustion, for instance, I might judge that giving a great deal to the poor out of feelings obligation to them is morally estimable, or judge that an employer's failure to help financially with an employee's medical expenses (motivated by a desire to keep such funds for himself) is morally disestimable. In such a state I might fail to feel much of anything – no impartial versions of gratitude or feeling screwed over – and no motivation to emulate or disemulate the donor or the employer.

Similarly, we might consider someone who has been brought up to think that the poor are poor because they deserve their state, that feeling obligated to help them is being

suckered, and it is mere weakness to show compassion for the poor. After reviewing the relevant empirical evidence and running through the relevant philosophical arguments he might change his mind entirely, and come to think that it is in fact morally estimable to help the poor out of feelings of obligation or care (as much as, say, observing loyalty to one's worthy friends at extreme personal cost). It quite possible (especially at first) that his judgments about the moral estimability of those who help the poor will be unaccompanied by anything that actually feels like "looking up" to such people or impartial versions of gratitude towards them, and may also be unaccompanied by any actual motivations to emulate what they do.

Or consider someone who has grown up thinking that giving to one's employees more than what they have contracted for interferes with the purity of market and could be motivated only by muddled communistic thinking. After becoming convinced by the relevant empirical evidence and philosophical arguments,[124] he might change his mind entirely, and come to think that it is in fact morally disestimable not to help needy employees with medical expenses out of motives to keep such funds for oneself. But again, it is possible (especially at first) that his judgments about the moral disestimability of employers who fail to help their employees will not be accompanied by anything that feels like "looking down on" or "feeling revolted by" such behavior, an impartial version of feeling screwed over by the employers, or motivations to disemulate their behavior.

Just like those who bloodlessly judge that they have done culpable wrong or who bloodlessly judge a state of affairs good, the above kinds of people can be making genuine judgments that the things in question are morally good or bad without having any actual moral esteem or disesteem towards them. The mere fact that they lack these attitudes does not mean that they are merely "paying lip service" to any moral views that they espouse. For their moral judgments may play their characteristic roles in inference, decision, and evaluation quite without their succeeding in giving rise to the esteem and disesteem the fittingness of which they entail.

[124] E.g. that the purity of the market is not so threatened, that one can care for or feel obligated towards employees while vigorously opposing the collective ownership of assets, and that the mere fact that one has contracted with a friend or person in need does not automatically relieve one of one's moral reasons to help her.

If someone thinks that giving money to the poor out of feelings of obligation is morally estimable, he may have tendencies to infer that similar kinds of actions are morally estimable, such as providing the poor with emergency medical assistance. He may be less inclined, for instance, to dismiss such action as helping people who aren't worth helping, and even if he has no actual esteem for giving money to the poor, he might well have esteem for providing medical assistance to the poor on account of his inferential tendency. In a similar way, one who thinks it disestimable not to help an impoverished employee with a medical expense might be inclined to infer that similar acts are disestimable, like refusing to try to find an employee lighter duties if he finds his current duties physically difficult. Even if one feels no actual disesteem for those who refuse to help with medical expenses, one might well feel disesteem for those who refuse to help find struggling employees find physically more appropriate work on account of judging it disestimable, where one would not have had this attitude had one failed to make the inference from one's judgment about the disestimability of refusing to help with medical expenses.

Similarly, judging something morally good or bad involves tendencies to infer things about one's own and other people's reasons for action, which can make a behavioral difference even when one feels no actual esteem or disesteem. Those who think that helping the poor is morally good will tend to infer that they and others have reason to offer such help. Subsequently, they may be moved to praise it, vote for policies that encourage it, and, by the kind of "sucking it up" pathway we have seen, their views about the appropriateness of esteem for helping the poor may cause them to intend and thus be moved to give to the poor themselves even though they lack actual feelings of esteem for doing so. In the same way, those who think it morally disestimable to refuse to help impoverished employees with medical expenses will tend to infer that they and others have reason to offer such help. Subsequently, they may be moved to criticize it, vote for policies that discourage such employer refusal, and, again by means of "sucking it up," form and be moved by intentions not to do things like this when in the relevant circumstances, even though they lack actual feelings of disesteem for doing so.

Finally, since judging something morally good or bad is a matter of judging it fitting to esteem or disesteem it, these judgments will involve tendencies to view one's

failure to have the relevant attitudes as problematic in a way that may comes out in other attitudes and behavior. One may tend to direct attention in ways geared towards instilling the attitude in question (by for instance reviewing the evidence in favor of the moral estimability or disestimability of the thing in question, attempting to imagine the relevant cases vividly, relating this information to more emotionally charged or personal experiences, etc.). One may feel badly (e.g. ashamed) about one's failure to have the attitude, possibly upbraid oneself public or private for the failing, and in the more extreme cases even go to see a therapist to attempt to do something about the defect or fact that something has gone wrong with one for failing to have the attitude (even independent of the pragmatic costs of failing to have it).

In much the same way, we can see that one can morally esteem something that one judges to be in no way morally good, and one can morally disesteem something that one judges to be in no way morally bad. Suppose that someone had been heavily acculturated by Nazi propaganda and had once believed that overcoming social criticism and feelings of squeamishness towards killing Jews in the service of Nordic dominance was highly morally estimable.[125] After reviewing the relevant philosophical arguments the person might change his mind entirely, and come to think that there is nothing at all morally admirable about such killings – that the dominance of Nordics over non-Nordics really isn't a morally worthy goal after all, that "squeamishness" towards harming Jews (perhaps constituted by feelings of care or obligation towards them) is in fact warranted, and that any social criticism of such harms would be entirely justified. Still, this person might (especially at first) still find his blood stirred by images of storm-troopers going against the softer sides of their nature and the complaints of respectable society in carrying out the final solution. He might, that is, have the same feelings of "looking up to" the storm-troopers, which feels moralistic in that it is reminiscent of something like an impartial version of gratitude that makes him feel like thanking them, he might fantasize or wishfully imagine being and doing like the storm-troopers, and might find himself with motivations to emulate the storm troopers and do likewise. He might have all of this, even as he judges his moral esteem for storm troopers unfitting, with its

[125] For discussion of how just this sort of thinking went on see Sabini and Silver (1982).

characteristic tendencies to inference, motivation to check his esteem, and evaluation of his esteem as defective.

Similarly, suppose that someone that had been heavily acculturated in the early 20th century deep south, and had once believed that Northerners who came south to agitate for black equality out of feelings of care for and obligation to blacks were very morally bad or disestimable. After reviewing the relevant philosophical arguments he might change his mind entirely, and come to think that there was nothing at all morally disestimable about Northerners doing this – Southern institutions were after all unjust, the feelings of care and obligation on the parts of Northerners were entirely appropriate, and the Northerners had legitimate reasons of other-defense to do what they did. Still, this person might still find himself peculiarly repelled by images of the Northerners coming South, perhaps with very little first-hand knowledge of Southern conditions, upsetting the very institutions that his confederate-soldier-turned-clan-member grandfather had fought for,…(one gets the idea). He might, that is, have the same feelings of "looking down on" the northerners, which feels moralistic in that it is reminiscent of feeling like they are jerks, or an impartial version of feeling screwed over by them. He might find himself fantasizing or wishfully imagining doing and being unlike the Northerners in similar circumstances, and he might find himself with motivations to disemulate the Northerners and do other than they if the occasion were to arise. This might be so, although he judges his moral disesteem unfitting, with characteristic tendencies to inference, motivation to oppose his disesteem, and tendencies to evaluation of his disesteem as defective.

At this point, we should be in a position to appreciate how the lessons of Chapter 2 apply to the present case. The view that moral esteem and disesteem involve judgments of moral goodness and badness entails that episodes of recalcitrant esteem involve conflicting moral judgments, but it seems that mere recalcitrant esteem involves none of the conflicting tendencies to inference, decision, or evaluation that would be involved in a conflict of moral judgment. The judgmentalist could "go-quasi" – maintaining that moral esteem and disesteem involve only sub-judgmental "evaluations" of moral goodness and badness. But moral esteem and disesteem can be recalcitrant, not only in the face of our judgments, but indeed in the face of our intuitions – we can imagine the

above characters suspending judgment about the warrant of their esteem and disesteem, but having nagging suspicions that their esteem for stormtroopers and their disesteem for northerners are unwarranted. On a version of quasi-judgmentalism where the moral evaluations involved in esteem and disesteem are cognitively penetrable perceptions or intuitions, we would have to have conflicts of intuitions wherever we have intuitively recalcitrant esteem and disesteem. Yet intuitively recalcitrant moral esteem and disesteem need not manifest such conflicts – all of the spontaneous tendencies to assent, inference, decision, and evaluation of the ex-Nazi and ex-Southern-racist could be on the side of intuitions that stormtroopers are not estimable and Northerners are not disestimable, and none on the side of their being moral good or bad respectively.

The judgmentalist could try retreating to a version where the "moral evaluations" involved in moral esteem and disesteem are cognitively impenetrable, but it looks metaphysically mysterious how there could be such representational contents, and in any event this would be to give up on the judgmentalist's explanation of the guidance of our moral esteem and disesteem by the judgments of goodness and badness they allegedly involve. But, come to think of it, that story couldn't explain how our judgments of goodness and badness directly govern the conative elements of esteem and disesteem anyway. The governance is imperfect, as the above cases illustrate, but is still quite genuine. For instance, changes in basic normative views about the moral significance of property rights can affect relatively frictionless changes in our attitudes of esteem and disesteem for those who may, at significant personal risk, either take from the rich to give to the poor out of concern for the poor, or defend the property of the rich from the depredations of the poor out of concern for the rich. In the same way, changes in basic view about the moral importance of autonomy can affect relatively smooth changes in our feelings of esteem for parents who devote a great deal of time and effort on interventions in the lives of their older children, or for parents who exercise great self-control in letting their children live as they see fit. These changes in view genuinely affect the kind of affect we have towards the relevant parties, the tendencies we have to wishfully imagine being like or unlike them, and the kinds of motivations we have to emulate or disemulate them. These changes are also instances of different normative views making a difference to conation in the absence of differences in views about what

our merely descriptive circumstances are like. So it's no-go for the kind of preestablished harmony account of why changes in moral esteem and disesteem correspond to changes in views about goodness and estimability solely in terms of both being sensitive to the same changes in empirical views about circumstances.

With all of that out of the way, the path is clear for us to understand moral esteem and disesteem as distinctive syndromes of affect, attentional focus, and motivation, which do not involve representations of moral goodness or badness. We can thus reassure ourselves that the above fitting attitude analyses of moral goodness and badness in terms of the fittingness of moral esteem and disesteem are non-circular and quite informative. In conjunction with our understanding of judgments about the fittingness of valenced attitudes as basic assessments of warrant with the same propensity to directly guide these attitudes as judgments about epistemic reasons for belief guide belief, our fitting attitude analyses can explain how our moral judgments have a tendency (though one that is by no means iron clad) to directly govern the syndromes of affect, attentional focus, and motivation that constitute our attitudes of moral esteem and disesteem.

4.2. Reasons to Be Good and Reasons Not to Be Bad

Given the understanding of the relationship between valenced attitudes and reasons for action that we developed in Chapter 3, the fitting attitude analyses of moral goodness and moral badness entail that we have reason to be morally good and to avoid being morally bad. The fitting attitude analyses of moral goodness and moral badness entail that the following are analytic truths:

(1) If an action, motive, character trait, or agent is morally good, then it is fitting to morally esteem that action, motive, trait, or agent.

(1') If an action, motive, character trait, or agent is morally bad, then it is fitting to morally disesteem that action, motive, trait, or agent.

Now, as we have seen, moral esteem involves motivation to emulate its object, and moral disesteem involves motivation to disemulate its object, or more precisely:

(2) To morally esteem an action, motive, character trait, or agent involves being motivated to perform that act, behave like someone who has that motive or trait, or behave like that agent, when in similar circumstances oneself.

(2') To morally disesteem an action, motive, character trait, or agent involves being motivated to omit that act, behave unlike someone who has that motive or trait, or behave unlike that agent, when in similar circumstances oneself.

Recall then from Chapter 3 our Warrant Composition Principle, which states that an attitude's fittingness entails the fittingness of its essential components. (2) and (2'), together with the Warrant Composition Principle thus entail:

(3) If it is fitting to morally esteem an action, motive, character trait, or agent, then it is fitting to be motivated to perform that act, behave like someone who has that motive or trait, or behave like that agent, when in similar circumstances oneself.

(3') If it is fitting to morally disesteem an action, motive, character trait, or agent, then it is fitting to be motivated to omit that act, behave unlike someone who has that motive or trait, or behave unlike that agent, when in similar circumstances oneself.

But as we saw in Chapter 3, we can explain how practical reasoning guides our motivations and actions by identifying considerations that contribute to motives' fittingness with reasons to perform the actions that realize their objects. This was summarized in the Motivation-Action Principle, which entails that:

(4) If it is fitting to be motivated to perform an act, behave like someone who has a motive or trait, or behave like some agent, when in certain circumstances, then one has

reason to perform that act, behave like a person who has that motive or trait, or behave like that agent, when in those circumstances.

(4') If it is fitting to be motivated to omit an act, behave unlike someone who has a motive or trait, or behave like unlike some agent, when in certain circumstances, then one has reason to omit that act, behave unlike a person who has that motive or trait, or behave unlike that agent, when in those circumstances.

And, of course, (1), (3), and (4), and (1'), (3') and (4') together entail:

(C) If an action, motive, character trait, or agent is morally good, then one has reason to perform that act, behave like a person who has that motive or trait, or behave like that agent, when in similar circumstances oneself.

(C') If an action, motive, character trait, or agent is morally bad, then one has reason to omit that act, behave unlike a person who has that motive or trait, or behave unlike that agent, when in similar circumstances oneself.

Importantly, the reasons spoken of in (C) and (C') can coexist along side other reasons to do things that are like those done by the morally good, and to omit things that are like those done by the morally bad, which have no essential connection to the fact that these acts resemble those of the good or the bad in these kinds of ways. We might, for instance, think that it is morally good to alleviate the suffering of the poor out of care for them or out of feelings of obligation to them. By (C), our view entails that we have reason to alleviate the suffering of the poor due to its being fitting to help them out of esteem for those who do so out of care or feelings of obligation. But it seems that we might respond to reasons to help the poor that are quite independent of the fact that doing so out of care or obligation is morally good. Think, for instance, of those morally good people who actually do help the poor out of care or feelings of obligation. Evidently, caring about someone intrinsic motivation to help her, or motivation to help her for her

own sake.[126] If this care is fitting, as care for the poor presumably is (as, it seems, is care for anyone), then we have again by the Warrant Composition and Motivation-Action Principles that one has reason to help the poor, here due to its being fitting to help them out of care for them. In the same way, it might well be fitting to feel obligated to help the poor – since, after all, here they are, suffering through no fault of their own, where you can do something to help, and they greatly need this help. And since as we have seen feeling obligated to help the poor involves motivation to help them, we have by WCP and MAP that one has reason to help the poor, here due to its being fitting to help them out of feelings of obligation to do so. So we might well, like the morally good, help the poor for the reasons we have due to the fittingness of doing so out of care or feelings of obligation to them, rather than for the reasons we have due to its being fitting to help them out of esteem for those who help them for the aforementioned reasons.

This might seem to remind us of a familiar feature of morally good conduct, namely that those who engage in it seem primarily concerned not about its character as morally good but rather about something else, like the people they help. Indeed, were they to be otherwise concerned, it would seem that their conduct would no longer be morally good, or at least not as morally good as it is.[127] Still, we seem to think that an act's character as morally good carries with it a certain kind of reason to perform it. What the fitting attitude analysis together with WCP and MAP may help us explain, however, is that it while such reasons to perform acts that are guaranteed by their moral goodness certainly exist, these are not the reasons to which morally good agents are actually responding. Those who respond to such reasons are primarily those of us who do not have the morally good motives, and can only emulate in conduct those who do.

But do the morally good themselves have the reasons guaranteed by their moral goodness to do what they do? I think that the answer delivered by the fitting attitude analysis of moral goodness together with WCP and MAP is 'yes'. A striking feature of esteem and emulation is that its object need not actually exist, or need not exist already. Before acting, a person can imagine an image of someone else or even her own future

[126] See Darwall (2002).

[127] Though of course it could happen to be just as morally good for other, serendipitous reasons. It might, for instance, be more difficult psychologically to steel oneself to help someone out of moral esteem for those who help her out of care, in such a way that one deserves a great deal of credit if one succeeds in doing so.

self doing something she is thinking about doing out of a particular motive or character trait – say giving money to the poor out of care for them. She can thus morally esteem this image, and with this esteem feel motivation to emulate or behave like it – i.e. feel motivated to give to the poor. If giving to the poor on the part of the imagined other person or her future self out of the care by which she imagines them to be motivated is in fact morally good, the fitting attitude analysis entails that it is now fitting for her to esteem these images, and since this esteem involves motivation to emulate them, the WCP and MAP entail that she has reason to emulate them and perform the act of giving to the poor. But, of course, if she responds to these reasons by feeling esteem and being motivated by emulation of the images, she may well turn out unlike the images in a crucial respect – she will be giving to the poor out of esteem for those who care, not actual care. And this may well not be (and probably is not very) morally estimable, or at least not as estimable, as the images she was (fittingly) esteeming and emulating.

We might take a similar look at the ways in which one can respond to different reasons to omit things that are like those done by the morally bad. An employer might, for instance, come to the conclusion that, while it wouldn't necessarily be wrong to help an employee, failing to do so out of sheer selfisheness would be "shabby" or morally bad. By (C), the employer's view entails that he has reason to omit this failure – that is that he has reason to help the employee due to its being fitting to help him out of disesteem for those who selfishly do not do so. But it seems that the employer might respond to reasons to help the employee that are quite independent of the fact that failing to do so is morally bad. We might, for instance, think that it is fitting to care about the employee or to feel obligated to help him (given his distress, the fact that one can help, etc.). If we are right, then by the Warrant Composition and Motivation-Action Principles the employer has reason to help the employee due to its being fitting to help him out of care for or feelings of obligation to him. Indeed, we might think that these latter kinds of reasons are really the most important. We might think that helping the employee simply in response to those reasons one has due to the fittingness of helping out of disesteem for those who do not help involves a kind of failure to attend to the most important things about the features of the case.

But while the reasons spoken of in (C) and (C') can thus coexist with other reasons to act like the morally good and avoid acting like the morally bad, we are not guaranteed that there will always be such other reasons. It might well be that a person could help others at great personal cost out of a kind of care that is so strong so as to be unfitting, or out of a kind of unfitting lack of self regard. Of course, if the care and lack of self concern are too crazy, we would seem to get something more like mental disease and less like moral goodness. But it seems coherent (and I think plausible) to think that there are some cases of actions out of unfitting motives that are still morally good. We might think in this regard of Christian love of one's enemies. Especially if those who mistreat one are genuinely culpable and vicious, it might be unfitting to care about them terribly much, and it might betray a kind of unfitting lack of regard for one's own interests to, say, "take the fall for" such a person and be punished in his stead. But it looks coherent to admit all this, and yet to think that selflessly taking the punishments deserved by those who mistreat one out of deep feelings of care for them can be morally estimable.[128]

In the same way we might think back to people who might feel guilt for breaking lamps through no fault of their own. We can agree that their feelings are unfitting, but think nonetheless that these feelings are morally estimable, or at least that the disposition to have them is a morally estimable trait of character. Should someone act out of such guilt and offer us a new lamp, we might well think of her as estimable on that account. We would probably be inclined to call the gesture "a nice thing to do," and this seems to expresses a thought that is very close (if not identical) to thinking of her gesture as morally estimable. Indeed, we might go further along these kinds of lines. We might think that certain failures to feel guilt even when it is completely unfitting (and perhaps also certain failures to act as one would who had such guilt) is itself morally bad or disestimable. Thus, suppose that a trolley driver slams on his breaks at the very instant a child pops into his view, but there is not enough time for the trolley to break. The driver, let us suppose, was exercising all due caution and did literally everything he could to prevent the death of the child. His feeling guilt for running over the child would thus be

[128] Though to make the line of thought plausible we might well have to rule out one's doing so out of the mindset of a victim in an abusive relationship.

unfitting. But it seems that we could fully agree with this, and still think it morally disestimable for the trolley driver to fail to feel such guilt, or, at least, that his disposition not to feel it would betray a morally bad cast of character. Moreover, were the driver to fail to do things like apologize to the child's parents and in other ways display the gestures and mannerisms of someone who feels guilt, we might well think his carrying on like he is just as blameless as a bystander is morally bad.

When in these kinds of cases a person does what is morally good or avoids doing what is morally bad out of motives that are quite unfitting, she may not be responding to anything that is a genuine reason for her to act. What is interesting, however, is that even in these kinds of cases we seem to want to say that there is a kind of justification for what the person did. The very fact (if it is a fact) that it is so good to take the fall for one's enemies seems to be capable of making it rationally permissible to do so, and the very fact that it would be so morally disestimable not to apologize, etc. for something that is not one's fault seems capable of making it the case that one has most and indeed conclusive reason to do these things. Our analysis so far reveals how there are in fact reasons that support doing these things (assuming they actually are morally good and bad respectively), which are entailed by the fact that the actions in question are morally good, or that failing to perform them would be morally bad. These are the reasons one has to do such things due to its being fitting to do them out of esteem for the morally good or out of disesteem for the morally bad. One thing that is interesting, however, is that in cases such as these it may be impossible to act in a way that is both morally good and responsive to genuine reasons. Thus, if we were to take the fall for those who mistreat us out of esteem for the ideal of selflessly doing so out of care, we would not ourselves be instantiating that estimable ideal. But if we were to instantiate that estimable ideal, we would be acting solely on motives that are unfitting and moreover unresponsive to what could possibly justify what we do – namely the reasons we have due to its being fitting to act from esteem for those act from care and a lack of self concern.

4.3. Moral Wrongness and Mandates for Feeling Obligated

We might say that an act (or omission) which is morally wrong not to perform is one which is *morally required*, or, in one good sense of the phrase 'moral obligation', we might say that an act which is morally wrong not to perform is *morally obligatory*.[129] So what we saw in Chapter 3 was that analyzing an act's moral wrongness in terms of the fittingness of feeling obligated not to perform it can, in conjunction with the Warrant Composition and Motivation-Action Principles, explain why an act's status as morally obligatory, or wrong not to perform, entails that we have reason to perform it. We have just seen how analyzing an action's moral goodness in terms of the fittingness of morally esteeming its performance can, in conjunction with the same principles, similarly explain how an action's goodness entails that we have reason to perform it.

But there might seem to be an important difference between the kinds of reasons guaranteed by an act's status as morally obligatory as opposed to its status as morally good. An action's moral goodness seems to entail that we have some reason to perform it, but not necessarily conclusive reason. It seems coherent (and indeed quite plausible) to think that there are morally good actions that it can be rationally permissible to fail to perform, like getting oneself shot in the head to spare a stranger the somewhat more painful death of drowning. Things seem different, however, with the kinds of reasons we have to perform acts that are morally obligatory, or morally wrong not to perform. There seem to be genuine problems with the coherence of thinking that one's doing something

---

[129] Perhaps one can also speak in English of 'moral obligations' in a "prima-facie" (or really a *pro tanto*) sense, or in a way in which one can have one "moral obligation" to ϕ without its being morally wrong for one to fail to ϕ if one has a "stronger moral obligation" not to ϕ. If so, then I merely wish to distinguish the concept I am expressing with 'moral obligation' from this other concept, and suggest that the concept expressed by 'moral obligation to ϕ' in this other sense is what I would call the notion of THERE BEING CONSIDERATIONS THAT CONTRIBUTE TO ONE'S BEING MORALLY OBLIGATED TO ϕ (OR THE MORAL WRONGNESS OF ONE'S FAILING TO ϕ).

I am told that there is also a sense of 'moral obligation' in English in which it would be true to say that it is morally wrong to let children drown or to fail to help the poor but that one is not morally obligated to help them. I do not have much of an idea of what it would be to speak of 'moral obligation' in this sense; perhaps it is something like to speak of what is "morally wrong not to do because of some contract or agreement you made." I really don't see why we would care about marking such a distinction. In any event I suspect that talk of moral obligation in this kind of more restrictive sense would have a pernicious tendency to make people speciously suspect that other things that are morally wrong not to do are somehow less important not to do because they are not 'morally obligatory'. So I emphatically will not speak of moral obligations in some more restrictive sense like this and I would encourage others not to do so either.

would be morally wrong but that one has most reason to do it anyway. As such, a strong version of what Stephen Darwall (1997, 306) calls the thesis of "morality-reasons internalism" might seem to be true of moral obligation, namely "if *A* is morally obligated to do *X*, then necessarily there is *conclusive* reason for *A* to do *X*."

It would be odd, however, if this strong thesis were to be explained solely in terms of the weightiness of the considerations that make acts morally obligatory or wrong. Such considerations – for instance that *I have promised to be across town* and that *she will die if I don't stay and help* – can be brought into conflict without necessarily giving rise to rational dilemmas[130], and it seems at least coherent to think that they are at times outweighed by non-moral reasons like *getting across town will get me killed*.[131] As such, I think that a much more attractive explanation of the strong thesis is that whether an act gets to count as falling under our concept of MORAL OBLIGATION – unlike, say, our concept of MORAL GOODNESS – is itself sensitive to whether or not the reasons in favor of performing it are actually conclusive. That is, as W.D. Falk (1948, 30-31) suggested, "our very thinking that we ought [that is, are morally obligated] to do some act already entails that, by comparison, we have a stronger reason in the circumstances for doing it than any other."

I think that the strategy I have been pursuing for analyzing moral concepts and explaining their connections to reasons for action can help explain why Falk's kind of account of the necessary conclusivity of reasons not to do moral wrong is in fact correct. The explanation will proceed in two steps. First, in this section, we will see how the sense in which it is fitting to feel obligated not to do what is morally wrong is that it is rationally mandatory to feel obligated not to do it, unless, that is, we are going to omit doing it anyway. In the next section we will see how rational mandates for feelings of obligation relate to what it is fitting to be moved to pursue on balance. We shall then put

[130] By which I mean situations in which whatever one does is rationally impermissible or such that one has conclusive reason not to do it. I, for the record, do not think that it is conceptually possible for there to be such situations.

[131] If the reader thinks that duties to oneself render this a moral reason, I invite her to consider whether there is some degree of trivialness of promise and some degree of harm that will befall one if one keeps it such that it is at least coherent to think that: (i) were it not for the harm to oneself one would be morally obligated to keep the promise, but (ii) given the harm one would incur by keeping the promise it is rationally permissible to break the it, yet (iii) one does not "owe it to oneself" to prevent the harm to oneself by breaking the promise.

these things together to see why we have conclusive reason not to do whatever is genuinely wrong.

Suppose that Bugsy and his unit are pinned down by a sniper, and that by chance a child is drowning in a pond a few yards in front of them. Bugsy knows, let us suppose, that the sniper will not shoot the child (they are of the same nationality), but that the sniper is also not above shooting anyone in his unit who goes to try to save her. Bugsy realizes that he would need only a few seconds to run over and pull the child out of the pond - enough time to save the child before the sniper can fire, but not enough time to make it back to cover before being shot.

In this situation, Bugsy might well feel obligated to save the child. If so, it seems plausible (or at least coherent) to think that his feeling of obligation is perfectly justified. Here is a child, terrified, drowning, desperately in need of help, and Bugsy knows that he can save her. Bugsy knows that this will cost him his life, but it seems justified for Bugsy to feel like he "just can't leave her" there to drown, even if the alternative is getting killed.

On the other hand, it might well be the case that Bugsy does not feel obligated to save the child. He might realize that helping the child will get him killed, and consequently not take seriously the option of running out and saving the child as something to do in his situation. He might feel horrified to have to watch the child drown, and curse the rottenness of the situation, but given his knowledge that emerging from cover will get him killed, he in no way feels any kind of prospective guilt-tinged aversion to staying put. Bugsy might care about children and feel obligated to do a good deal to keep them from harm, but sacrificing everything for the sake of a single child who he never met might be something that Bugsy does not feel he has to do. If this is how Bugsy feels, it seems plausible (or at least coherent) to think that his feelings would be justified in this eventuality too.

Indeed, it seems perfectly coherent to think both that Bugsy would be justified in feeling obligated to save the child, and that he would be justified in having no feeling of obligation to save the child. Thinking that Bugsy would be justified in not feeling obligated in no way commits us to thinking that any feelings of obligation Bugsy might

have would have to be unjustified. We would not have to think, that is, that his feelings of obligation are at most estimable, or disestimable not to have, like the guilt-feelings of a trolley driver who knows he did everything he could not to run over a child. We could well think the fact that the child in front of him desperately needs help and that Bugsy can provide it contributes to the fittingness of feeling obligated, and that the fact that Bugsy will be killed if he helps is insufficient to make it unfitting for him feel this way. But we could at the same time think Bugsy's circumstances to be such that in them it is in no way unfitting to lack feelings of obligation to help the child. We could, that is, have a case much like those I discussed in Chapter 3 under the heading of rationally optional attitudes. Just as Bugsy would be justified in feeling anger at those who have culpably wronged him, but allowed not to feel such anger, Bugsy could be justified in feeling obligated to save the child, but perfectly well allowed not to feel this way too.

If, however, we think in this way that it is merely rationally optional for Bugsy to feel obligated to save the child, we will not, it seems, think that Bugsy is morally required or morally obligated to do so. For consider a situation in which we did think that Bugsy was morally obligated to save a child. Suppose that instead of Bugsy in uniform with a sniper ready to kill him if he emerges to pull a drowning child from a pond, we merely had the case presented by Singer (1972), where Bugsy is a civilian walking past a shallow pond when he sees a child drowning in it, and all it will cost him to save the child is a few moments of his time and the muddying of his clothes. Here, we might think, Bugsy has the same reasons to feel obligated to save the child – that the child is in need of help and Bugsy can provide it – which make it fitting for Bugsy to feel obligated to save the child. But, in this kind of case, it does not seem fitting for Bugsy to lack feelings of obligation in the way he might justifiably lack them in the case with the sniper. It does not seem fitting for Bugsy to simply walk past the pond and be unmoved by any feeling of obligation to wade in, in the way it might be fitting for Bugsy to simply sit tight and be unmoved by any feeling of obligation to break cover in the case with the sniper. In Singer's case, it does not seem fitting, for instance, for Bugsy to be such that he would signal someone else that the child is drowning if someone else were around, but to have no feeling of obligation to pull out the child at the cost of muddy clothes to himself.

Of course, in Singer's case, there is quite another way in which Bugsy might not feel obligated to save the child. He might simply wade in and pull out the child out of natural feelings of care for her. Or he might wade in and pull out the child simply because he does not want to have to witness the child drowning, or simply because he does not want to experience the guilt that he knows he will feel for failing to help her. Or he might wade in and save the child out of fear that others would find out that he did nothing to save her and shun him as a result. If Bugsy is such that he is going to save the child out of these kinds of motives anyway, then there might well seem to be nothing unfitting about his failing to feel obligated to save her. But this is different from the way in which it is not unfitting for Bugsy to have no feeling of obligation in the case with the sniper. For there, it seemed, we could well think that it is fitting for Bugsy to have no feeling of obligation to save the child, even though he is not going to save the child anyway out of some other kind of motive.

The difference, then, between Singer's case, where (let us assume) we think that Bugsy is morally required to save the child, and the case with the sniper where (let us assume) we think Bugsy is not so required, seems to be this. In thinking it morally permissible for Bugsy to fail to save the child in the case with the sniper, we think that Bugsy is justified in having no feeling of obligation to save the child even if he is not going to save the child anyway. But in thinking it morally wrong for Bugsy to fail to save the child in Singer's case, we think that, if Bugsy is not going to save the child anyway, then it is unfitting of him to fail to feel obligated to do so.

This relationship between moral requirements and rational mandates for feelings of obligation in the absence of sufficient motivation to do what is morally required seems to hold quite generally. There are, of course, many things that it would be morally wrong for us to do that would it would not even cross our minds to do. When this is so, we might be quite permitted in having no feeling of obligation not to do them – not even, that is, a feeling of obligation in the dispositional sense. These things, like massacring everyone in our household with a machete, or pushing our best friend out of a window, might be so far from our inclinations and thoughts that we have no meaningful disposition to feel obligated not to do them (though we might acquire such a disposition were our inclinations to change). And this is surely perfectly fitting. But it seems to be

part and parcel of thinking something morally wrong to think that, in circumstances in which we were not already sufficiently deterred from doing it, it would be fitting to feel obligated not to do it, and moreover unfitting not to feel this way.

With this in mind, we might reflect back on the considerations in favor of the analysis of moral wrongness in terms of fitting feelings of obligation to see what exactly these considerations support. Suppose we lived in a land (which might well be our actual land) where very late term abortions were so dangerous and so heavily legally sanctioned that none of us would dream of having a very late term abortion of a perfectly healthy fetus. If some of those of us who are in the late terms of pregnancy became convinced by philosophical arguments that it would be deeply morally wrong to abort our healthy fetuses in our actual circumstances (where, let us suppose, we did not believe this before), it is not as though we would all of a sudden feel strongly obligated not to abort them. For we would not be the least bit tempted to do so anyway. What seems to be guided by philosophical inquiry into wrongness, and the judgments about wrongness that ensue, are feelings of obligation not to do things that we are not already deterred from doing, as was the case of our person who became convinced to be vegan in Chapter 3. If that person had already been sufficiently deterred from eating meat, dairy, and eggs because, say, she was deathly allergic to all that stuff, she too would presumably acquire no feelings of obligation to avoid eating these things upon being convinced of her moral obligations not to do so. Of course, our judgments of wrongness don't perfectly guide our feelings of obligation even when we are not sufficiently deterred from doing the things we come to judge to be wrongful – our lack of feeling obligated can in this kind of way be recalcitrant to our judgments. But the point here is that, if one is already sufficiently deterred from doing something that one judges to be morally wrong, one's failure to develop a feeling of obligation to doing it is no evidence of recalcitrance; it is not an instance of our feelings failing to align with our views about what they ought to be.

In the same way, our judgments about wrongness seem to have more of an effect on our feelings of obligation than simply judgments about the rationally optional nature of feeling obligated in certain circumstances. Consider, first, the difference between

coming to judge that an attitude is optional and coming to judge that an attitude is mandatory. If you come to judge that an attitude is rationally mandatory – say a desire for late term fetuses to survive, or a credence above .5 that the many worlds interpretation of quantum mechanics is true – then you will tend to have that attitude if you do not currently have it. Moreover, if you fail to come to have that attitude, your attitudes will be recalcitrant to your judgments – you will be failing to feel as you think you should, along with the kinds of negative evaluations of your response and cognitive dissonance that usually ensues. But if you come to judge that an attitude – say a desire to play Go, or a feeling of anger at someone who has harmed you – is merely rational optional, you may well exhibit no tendency to have this attitude if you do not already. What is more, your failure will not be an instance of your attitudes being recalcitrant to your judgments. You can think now that anger or desires to play Go would be justified (where before you thought that the person had no choice, or that Go is a silly, stupid game), yet have no actual anger or inclination to play Go, without the least bit of recalcitrance. Your attitudes are not somehow failing to respond to what you think they should be, you need have no negative evaluations of your failure to so respond, and you need be in no state of cognitive dissonance. The most immediate upshot of judging an attitude rationally optional seems rather to be to protect our ability to continue to have it in a state of non-recalcitrance if we should happen to have it already.

To see this in the case of feelings of obligation, suppose that Bugsy from our sniper case is the second way we considered he might be: he has no feeling of obligation to save the child, given that he knows that saving the child will cost him his life. But suppose we got hold of him and convinced him that it would be rationally optional for him to be the first way that we considered he might be. "Look, here, Bugsy, imagine you did feel obligated to save the child on account of the fact that she desperately needs help, and that you can help her. Would your feelings of obligation not be justified?" If Bugsy were sufficiently philosophically minded, it seems that he could agree without having to change his mind about his own justification for not feeling obligated that he would also be quite justified if he were to feel obligated. "Yes," Bugsy might reply, "I would then be as justified in my feelings as I am now – for in this situation there is more than one way that a person may justifiably feel." Bugsy's acquiring this view, it seems, could

exert no pressure in the direction of altering his lack of feeling obligated to save the child, and his remaining as is, with no feeling of obligation, would in no way be recalcitrant. Bugsy need be party to no cognitive dissonance and no negative evaluations of his responses.

But judging that it is morally wrong to do something has a different kind of effect on your feelings of obligation in cases in which you are not already sufficiently deterred from doing the thing you judge wrongful. If you are not already sufficiently motivated to give a certain amount of help to the poor, then coming to judge that it would be wrong not to do so has a propensity to make you feel obligated to do so. If you are not already deterred from doing things that you know contribute to the suffering and death of non-human animals, then coming to judge that you are morally obligated to do so has a propensity to cause you to feel obligated not to do this. If you are not already sufficiently deterred from aborting a late term fetus with Downs Syndrome, and are seriously thinking about doing it, then coming to judge that it would be wrong to do it has a propensity to cause you to feel obligated not to do it. What is more, if in any of these situations you continue to be insufficiently motivated to refrain from doing the things you judge wrongful, and yet you fail to feel obligated not to do them, your absence of attitude will be recalcitrant to your judgment. You will be failing to feel the way you think you ought, and this will involve a tendency to view your lack of response as something defective, to experience a certain kind of cognitive dissonance, and so on.

Thus, a judgment that it is morally wrong to do something governs feelings of obligation not to do it in the manner of a judgment that it is rationally mandatory to feel obligated not to do it in circumstances in which one is insufficiently deterred from doing it. The identification of judgments about wrongness with this kind of fittingness assessment in particular is thus what is ultimately supported by the arguments of Chapters 2 and 3 that some such fitting attitude analysis can explain the semantic, epistemic, attitude-guiding, and practical features of judgments about wrongness. In light of this, we might clarify the fitting attitude analysis of moral wrongness that we presented in Chapter 2 along the following lines:

Fitting Attitude Analysis of Moral Wrongness*:

To judge that agent *A*'s act of ϕ-ing is morally wrong is to judge that, unless *A* is already going to refrain from ϕ-ing anyway, it is rationally mandatory for *A* to feel obligated not to ϕ (or equivalently: to judge that, unless *A* is already going to refrain from ϕ-ing, it is rationally mandatory for *A* to feel prospective guilt-tinged aversion towards ϕ-ing).

4.4. Mandates for Feeling Obligated and States of Overall Motivation

Since feeling obligated not to do something involves being motivated not to do it, our refined understanding of the fitting attitude analysis of moral wrongness will, in conjunction with the Warrant Composition Principle, entail that if something is genuinely morally wrong (and one is not going to omit it anyway), then it is rationally mandatory to be motivated not to do it. Now in Chapter 3 we saw that rationally mandatory motives can, unlike rationally optional motives, entail the existence of conclusive reasons for action. But the mere fact that it is rationally mandatory to be motivated to do something does not, by itself, entail that one has conclusive reason to do it. For it could be that it is rationally permissible, or even rationally mandatory, to be more strongly motivated to do something else.

One example of this, which we saw in Chapter 3, was that of being stuck equidistant between two drowning swimmers: Bugsy and Suzy. We might think, for instance, that it is rationally mandatory to care for both of them, and to want neither to die under the circumstances. As such, we might think that it is rationally mandatory to want to save Bugsy, and also rationally mandatory to want to save Suzy. Since wanting to save someone involves motivation to save her, this entails, by the Warrant Composition Principle, that it's rationally mandatory to be motivated to save Bugsy and to be motivated to save Suzy. Of course you cannot save both, but it might seem in such a case that it is rationally mandatory to have equally strong, conflicting preferences. Since you must save someone, you must form an intention to save one of them and act on

it. But whichever act you perform, we might think it mandatory to retain a weaker, but still present, desire to perform the other. The mere fact that you cannot save both does not mean that you should not want to do so.

Or consider another kind of example that I mentioned in Chapter 1. We might well think, for instance, that it is rationally mandatory to want to keep your leg. Having a leg is ever so useful, and we might well be imagining a circumstance in which high quality prosthetic legs are not widely available. Similarly, we would think that (having a future well worth living) it is rationally mandatory to want to keep your live; failure to have this desire would constitute an irrational state like suicidal depression, no doubt. But suppose you face a case like that discussed by Kagan (1998, 86) where you are trapped under a tree, and will die unless you order someone to cut off your leg to free you. In such a case, we might think, it is still rationally mandatory to want to keep both your leg, but it is also rationally mandatory to have an even stronger desire to keep your life. "Live hard, die young, and leave a good-looking corpse" might express a coherent state of desire, but it is surely seems to be an irrational state of desire. You must, we might think, care more for your life than your leg. But we might well think that the mere fact that you must lose your leg if you are to keep your life does not count against wanting to keep your leg at all. That you cannot keep both leg and life does not mean that you should not want to keep both.

But now, it seems, we may have a difference between rational mandates for attitudes like desires and rational mandates for feelings of obligation. For suppose that you have made a promise to John to deposit his money before the bank closes, and closing time is near. The fact that you have promised this to John (and that it is a decently important promise by the way) is a significant reason, no doubt, to feel obligated to get yourself to the bank before closing time if you find yourself tempted not to do so. If you found yourself inclined to stay in to watch a movie, or feeling too tired to drive over to the bank, it would, it seems, be rationally mandatory to feel obligated to get to the bank. But suppose that on your way over to the bank you see Mary lying on the ground, hemorrhaging blood. There is no one else around and you realize that she will die if you don't stay and help. Of course, you also realize that if you stay and help the bank will close and John's money won't get deposited. But here, it seems, that the fact that you

cannot make it to the bank if you are to help Mary *is* a reason against having to feel obligated to make it to the bank.

Similarly, suppose that, in a case like that discussed by Foot (1967) and Scanlon (2000), you are a medic who comes across someone with an advanced case of disease *X*, which can be cured only with your entire supply of drug *M*. Drug *M* is somewhat valuable, and you were rather planning to sell your supply to buy something nice, so you find yourself a bit tempted not to offer to help the man with your supply of the drug. But here is this man who will die if you don't help, and that something nice is, as some like to say, of "no moral significance." So, we may suppose, given your temptation not to help, it is rationally mandatory for you to feel obligated to save the man's life by giving him your drugs. Let us suppose, however, that just as you are about to offer him your supply of drug *M*, you see that there are five other people, each of whom are also suffering from disease *X*. These five others are, however, less far along than the first, and each need only one fifth of your supply of drug *M* to survive. In such a case, it seems, the fact that you cannot save the one if you are to save the five counts against having to feel obligated to save the one.[132] Of course, as in the case with Bugsy and Suzy, it might be rationally mandatory for you to want all of these people to live, and thus want to save the one as well as the five. But, given that you cannot save them all, it seems that you no longer have to feel obligated to save the one.

Now as we have seen, the fact that it is not rationally mandatory to feel obligated to do something does not mean that it is rationally mandatory that you *not* feel that way. In the case with drug *M*, you might well continue to feel obligated to give the drug to the one even after you see that there are five others who need it. For here is this man who needs your help, and you can help him. We might well think that this would continue to justify your feeling like you just can't leave him to die even when the five others come on the scene. Our sense that, given that you cannot save everyone, you do not *have* to feel obligated to save the one, is just that. Because another way you could be in the case with drug *M* is such that, given that five equally deserving others need the drug too, giving the

[132] If, by the way, you are what some of us would regard as absolutely mad, let us suppose that you have flipped your coin or rolled your six-sided-die or what have you and it has come up in favor of saving the five (people like Taurek (1977), you know who you are). But even before you do this, it would seem that you no longer have to feel obligated to save the one. After the introduction of the five, that will happen, if at all, only if your randomization procedure favors him.

whole drug to the one is not something that you feel you have to do. And this, we see, is a perfectly fitting way to feel, even if it would also be fitting to continue to feel like you have to help everyone.

(Interestingly, though, there do seem to be cases in which the factors that remove the rational mandate for feeling obligated to do something also remove the rational option to feel obligated to do it. This might be so in our case where the fact that you must stay to save Mary's life removes the mandate to feel obligated to get to the bank. Given that Mary will die if you don't stay and help, it not only seems permissible to feel no obligation to get across town, it actually seems downright mandatory.)

It looks, then, that while it can be rationally mandatory to continue to want to do something even when you must be more strongly motivated to do something else, it cannot be rationally mandatory to continue to feel obligated to do what you may be most strongly motivated not to do. In our cases of saving Mary and drug *M*, you start out with reasons to feel obligated to keep your promise and to save the one man, which initially support being most strongly motivated to keep your promise and to save the man. Circumstances then change, and you get reasons to be at least as strongly motivated to do something else. But when this happens, your reasons to feel obligated to do the initial thing are no longer strong enough to make it rationally mandatory to have *any* feeling of obligation to do it at all. This is in marked contrast to how things stood with reasons for desire in the Bugsy-Suzy case and the leg-life case. There, you have reasons to desire to save Bugsy and to desire to keep your leg (e.g. *Bugsy will be connected to a life well worth living* and *A leg is an ever so useful thing*). You have, however, reasons to be at least as strongly motivated to do something else (save Suzy and save your life). Yet this in no way prevents your reasons to desire to save Bugsy and to desire to keep your leg from making it rationally mandatory to have these desires to some extent; it merely prevents it from being rationally mandatory to be most strongly inclined in their direction.

It would appear, then, that there is something about feelings of obligation that makes them rather different from other motivational states like desires. This is that it can only be rationally mandatory to have any feeling of obligation to do something so long as it remains the case that one is unjustified in being in at least as strongly motivated to do

something else – only so long, that is, as it remains rationally mandatory to be most strongly motivated to do what it is a feeling of obligation to do. It seems, then, that rational mandates for feelings of obligation "contour to fit" other kinds of considerations, in that they fail to hold if other considerations win out over the considerations in favor of feeling of obligated in the determination of what one should be most strongly motivated to do. We might thus refer to this feature of feelings of obligation as the:

> **Contour Thesis**:
> If it is rationally mandatory for agent *A* to feel obligated to do *X*, then it is rationally mandatory for *A* to be most strongly motivated to do *X*.

Now one might think that cases like the following are counter-examples to the Contour Thesis. To borrow (and use for slightly different purposes) an example from D'Arms and Jacobson (1994, 742-743), suppose that your mother has always deeply feared being put in a nursing home. Unfortunately, you are finding yourself unable to care for her very well and things are becoming a strain on your children. In such a case, we might well think it fitting for you to be most strongly motivated to put your mother in a home. But it might seem consistent with this to think that it is rationally mandatory to feel obligated not to do so. It might, for instance, seem to be consistent with thinking that you should be moved on balance to put your mother in the home for you to think that there would be something wrong with you if you failed to feel obligated not to do so. Similarly, it might seem consistent with your thinking that you should be moved on balance to put her in the home that to think that it would be inappropriate or unfitting for you to feel no kind of reluctance towards putting her in the home, or to be able to put her in the home "with perfect equanimity."[133]

One thing we can agree on from the start is that it is consistent with your thought that you should be moved on balance to put mother in the home for you to think that it would be rationally permissible for you to feel obligated not to do so. That is, it is consistent with your views to think that it would be permissible to feel obligated not to put your mother in the home, just as it is permissible to feel obligated to save the child in

[133] My thanks to Stephen Darwall for this last way of putting the intuition.

the case with the sniper or to give the drug to the one in the case with drug *M*. This is no threat to the contour thesis, which concerns itself only with rational mandates for feelings of obligation. Our question, then, is whether it is consistent with your views to think it not just rationally permissible but moreover rationally mandatory to feel obligated not to put your mother in the home.

The first reason we are given to think this stronger thesis consistent with your views is our intuition that it is consistent with your views to think "that *there would be something wrong with you* if you failed to feel obligated not to put your mother in the home." But, to the extent that this is really consistent with your views, I suspect that it is a thought to the effect that it would be morally bad or disestimable of you to fail to feel obligated not to put your mother in the home. Recall again our trolley driver, for whom we agree guilt would be unfitting but for whom it would be morally bad or disestimable not to feel. We might well put our intuition here as a view to the effect that "there would be something wrong with the trolley driver if he were to fail to feel guilt." There might be other cases too in which it might be rationally permissible to feel obligated to do something, and while it would not be unfitting to fail have this feeling of obligation, it would be morally bad or disestimable not to. We might well, for instance, imagine someone thinking this sort of thing about our case of the drowning child and the sniper. Someone might think that the second way Bugsy could be – such that he has no feeling of obligation to save the child on account of the fact that doing so would cost him his life – is fitting or justified but perhaps somewhat disestimable.

Indeed, from what we have seen it might be natural to confuse thoughts about the disestimability of having an attitude with thoughts about its unfittingness. For if having not feeling of obligation not to put mother in the home really is disestimable, it follows from our analysis of disestimability, the fact that disesteem involves motivation to disemulate, and the Warrant Composition Principle that one has reason to be motivated to do what one would do if one felt obligated not to put mother in the home – reason, that is, to be motivated to omit putting her in the home. Indeed, if moral disesteem for not feeling obligated to put mother in the home is rationally mandatory, it will be rationally mandatory to have these motivations not to put her in the home. But we can all agree that rational mandates for moral disesteem, like rational mandates for desires, are insensitive

to whether it is rationally mandatory to be motivated on balance to do what the disesteem motivates you to do. The Contour Thesis is concerned only with rational mandates for feelings of obligation proper, not rational mandates for disesteeming failures to feel obligated.

Another reason it would be natural to confuse judgments about an attitude's disestimability with judgments about its unfittingness is that, for reasons discussed by Velleman (2002) the former kinds of judgments can have effects on the attitude in question that are similar to assessments of its unfittingness. We shall have more to say about this in chapter 6, but the basic story is this. As we have seen, disesteem for an attitude involves tendencies to fantasize about and wishfully imagine not having it in the relevant circumstances. There seems to be a psychic causal pathway by means of which the right kinds of such simulation and play acting can give rise to the real McCoy – namely absence of the attitude in the relevant circumstances. This pathway is, moreover, not intentionally induced; over some time, it comes about that one lacks the attitude without one's having to do anything to bring it about. In bringing out the intuitive features of fittingness assessments, I have implicitly relied upon differences between the way fittingness assessments regulate attitudes without our having to do anything and the way this other kind of thing works – for instance, that fittingness assessments can often work their influence immediately, that they are generated through basic normative inquiry, and that, to the extent that judgments of an attitude's estimability or disestimability normatively regulate the attitude, they do so in a way that is parasitic on fittingness assessments of esteem and disesteem. To discuss these things further would be to get ahead of ourselves; we shall take them up in chapter 6 where we shall argue that the causal manifestations of fittingness assessments support our identifying them with a particular kind of functional state. For our purposes here we can simply note that judging an attitude unfitting and judging an attitude disestimable are similar in that they do not require us to do things in order to bring it about that we do not have the attitude, and that we should thus not be surprised if people occasionally confuse the latter with the former.

I see, then, no reason why we should not account for the intuition of the consistency between your view that you should be moved on balance to put your mother in the home, and your view that "there would be something wrong with you if you didn't

feel obligated not to put her in the home" by recognizing the latter as a judgment to the effect that it would be disestimable of you not to feel obligated not to put your mother in the home. Let us turn, then, to the second thought offered in support of the view that it is consistent with your views to think that it is rationally mandatory to feel obligated not to put your mother in the home. This was the intuition that it is consistent with your views to think it genuinely unfitting to have no kind of reluctance to putting your mother in the home, or to be able to put her in the home "with perfect equanimity." We have already seen two ways of accounting for the second kind of thought, consistent with its being something other than a thought to the effect that it is unfitting for you not to feel obligated to refrain from putting your mother in the home. The first we saw in our discussion of the case with drug *M*. For there, since you could not save both the one and the five, you did not have to feel obligated to give your drugs to the one. But we noted that it may be mandatory for you to care about all six, and –like in the case with Bugsy and Suzy – to want to save the one out of care for him. Your care for him should thus make you reluctant to withhold the drug, for you know that without it this poor man is going to die. Yet it was consistent with this that, since you cannot save them all, you do not have to feel obligated not to turn the one down. Much the same could thus be said about the case of putting your mother in the home. For surely, you will recognize, you absolutely should care about your mother and her aversion to being put in the home. This care should make you very reluctant to go against her wishes in this way. But, as in the case with drug *M*, this rationally mandatory reluctance is quite distinct from the feeling of reluctance constituted by feeling obligated not to put your mother in the home, which may, as you cannot care for her, the children, and so on, be something you do not have to feel.

A second way of accounting for the intuition that it is consistent with your views to think that you must feel reluctant to putting your mother in the home came up in our discussion of how you may think that it would be disestimable of you not to feel obligated to refrain from putting your mother in the home. As we saw, if this disesteem is fitting, so too is the motivation not to put mother in the home that it involves, and if this disesteem is rationally mandatory, then so too is motivation not to put your mother in the home. So if you think it disestimable for you not to feel obligated to refrain from

putting your mother in the home, and moreover that it is rationally mandatory to feel such disesteem towards the idea of your failing to have this feeling of obligation, your view will entail that it is rationally mandatory to be motivated not to put your mother in the home. So here again, we have rationally mandatory reluctance to putting your mother in the home that is not constituted by a rationally mandatory feeling of obligation to avoid doing so.

But it may also be worth considering some other ways in which it may be consistent with your view that you should be moved on balance to put your mother in the home for you to think it unfitting not to be reluctant to do so, which kinds of reluctance might feel more like feelings of obligation than reluctance out of care for your mother or reluctance out of disesteem for the idea of not feeling obligated to refrain from putting her in the home. The first is that you might realize that you are in a very difficult situation, and that it would be very easy to make the wrong choice. Given the stakes for those around you, you might think it morally obligatory to think very carefully about the matter, and to re-open the matter for deliberation if there is any hint that you are making the wrong choice. But, of course, the various considerations are constantly tempting you to decide one way or the other, or to conclude deliberation in their favor. So you might thus think that it is rationally mandatory for you to feel obligated to keep thinking carefully, keep checking your reasoning, and re-open the matter for deliberation in the face of hints that you are making the wrong decision. But this is distinct from thinking it rationally mandatory to feel obligated not to put your mother in the home if you conclude in favor of that option.

In a way that is perhaps related, it may well be that we are capable of having a certain kind of aversion to going against someone's interests or wishes, which resembles feeling obligated not to do so in certain respects, but which remains distinct from such feelings of obligation. We might call this attitude 'compunction'. Phenomenally, feelings of compunction might seem, like feelings of obligation, to be "guilt tinged" in a sense, but feeling compunction towards performing an act would involve something more like a feeling of hesitancy about, being unsettled about, or reluctance about performing it. The feelings of obligation not to do something or guilt-tinged aversion towards doing it in terms of which I have argued we can understand judgments of moral

wrongness do not seem so aptly characterized in these ways. As we have seen, they seem to involve something more like a feeling that one "just can't" or "just can't bring oneself" to perform the act towards which they are felt. Compunction towards doing something also seems to be more closely associated with the above mentioned kind of going back and forth about doing it or checking and re-checking to make sure that one's doing it would not be wrong to do. While one can of course feel obligated not to do something and be unsettled about whether or not to do it, feeling obligated not to do something seems in some sense to be more closely associated with being settled against doing it rather than simply feeling hesitant towards doing it.

We can, then, account for your thoughts that "there would be something wrong with you" if you did not feel obligated to refrain from putting your mother in the home, and that it would be unfitting for you not to be reluctant to do so, as something other than thoughts that it is rationally mandatory to feel obligated not to put her in the home. Given that you think that you should be motivated on balance to put mother in the home, these ways of accounting for the intuition that these other thoughts are consistent with your view seem at least as initially plausible as the claim that it is consistent with your view to think that it is rationally mandatory to feel obligated not to put your mother in the home. The (at least) equal plausibility of these ways of accounting for our intuitions of coherence thus neutralizes any force these intuitions might seem to have against the Contour Thesis, enabling the foregoing intuitions, about such cases as that of having to save Mary's life and that of the six who need drug *M*, to punch through and count decisively in favor of it.

4.5. Conclusive Reasons Not to Do Moral Wrong

We are now in a position to put things together to see why an act's moral wrongness entails the existence of conclusive reason not to perform it, thus vindicating the strong morality-reasons internalism thesis. We have, by our clarified analysis of moral wrongness from section 4.3, that an act's moral wrongness entails the existence of a rational mandate to feel obligated not to perform it. We then have, by the Contour Thesis

of section 4.4, that a rational mandate for feeling obligated not to do something entails the existence of a rational mandate for being most strongly motivated not to do it. Finally, we have by the Most-Motivation-Action Principle of Chapter 3 that a rational mandate for being most strongly motivated not to do something entails that one has conclusive reason not to do it. Hence, the entailment between an act's wrongness and conclusive reason not to perform it.

But to see more precisely what is going on, we must recall the exact nature of the rational mandate for feeling obligated guaranteed by an act's wrongness. For this mandate is not always in force – it is in force only when you are not already sufficiently deterred from performing the wrongful action. So to consider what follows from an act's moral wrongness, we must consider two general sub-cases: (C1) that in which you are *not* already sufficiently deterred from performing the act without having to feel obligated not to perform it, and (C2) that in which you *are* already sufficiently so deterred. By the above reasoning, we have it that if you are not already sufficiently deterred from performing the wrongful action, a rational mandate for feeling obligated not to perform it is in force, and it thus follows from the Contour Thesis that it is rationally mandatory for you to be most strongly motivated not to perform the action.

But consider now the other sub-case in which you are already sufficiently deterred from performing the wrongful action. There are here two sub-cases of this case: (C21) the one in which your state of overall motivation is fitting or justified, and (C22) the one in which your state of overall motivation is unfitting or unjustified. Consider first (C21), in which your state of sufficient deterrence from performing the wrongful act is justified. If our example of a wrongful action is beating innocents, then one way you might instantiate this sub-case is for you to care about innocents so much that you would never dream of beating them. Now there are in fact two sub-cases of this kind of justified sufficient deterrence as well: (C211) the case in which your state of sufficient deterrence is rationally mandatory, and (C212) the state in which your state of sufficient deterrence is rationally optional. As an example of (C211), we might think that it is not only permissible but moreover rationally mandatory for you to care about your loved ones so much that you would never dream of beating them. If your state of sufficient deterrence to performing the wrongful action is rationally mandatory, then *ex hypothesi* it is

rationally mandatory for you to be most strongly motivated not to perform the wrongful action.

But consider now the other sub-case (C212) in which your state of sufficient deterrence is rationally optional. Perhaps you would be rationally justified in caring about strangers so much that you would never dream of beating them no matter how much they annoy you, but perhaps this is not rationally mandatory to care about them this much. If your state of sufficient deterrence is rationally optional, then there will be other rationally permissible alternatives to it. For each of them, there are again two sub-cases as to what it will be like: (C2121) the one in which it is *not* a state of being sufficiently deterred from performing the wrongful act without your having to feel obligated not to perform it, and (C2122) the case in which it *is* a state of sufficient deterrence from performing the wrongful act in the absence of your feeling obligated not to. Thus, as a possible instance of (C2121), one permissible alternative to loving everyone so much that you'd never dream of beating them might be your being such that you are actually tempted, from time to time, to beat them. But given that doing *X* is morally wrong, we have from our clarified analysis of wrongness and our Contour Thesis that if you are insufficiently deterred from doing *X*, it is rationally mandatory for you to feel obligated to do *X*, and rationally mandatory for you to be most strongly motivated not to do *X* as a result. What this means is that your state of being insufficiently deterred absent feelings of obligation will only be rationally permissible if it also includes feeling obligated not to do *X* and being most strongly motivated not to do *X* as a result. So to be a genuinely permissible alternative to your loving everyone so much that you'd never dream of beating them, your state of not loving them so much needs to include feelings of obligation not to beat them and needs to be a state of being most strongly motivated not to beat them as a result.

Now, another rationally permissible alternative to caring about everyone so much that you would never dream of beating them might be one in which you do not care about them so much, but in which you go around in such a Zen-like state that the annoying things they do never bother you, and you are again never tempted to beat them. Here, then, we might have an instance of (C2122), where your alternative overall state of motivation is one of sufficient deterrence from performing the wrongful action. Of

course, since such a permissible alternative state is one of being sufficiently deterred from performing the wrongful act, it is, *ex hypothesi* a state of being most strongly motivated not to perform that act. So putting together our observations about (C2121) and (C2122), we can see that, if you are in a rationally optional state of sufficient deterrence to performing the wrongful act, your state will have rationally permissible alternatives, but each of those alternatives will have to involve being most strongly motivated not to perform the wrongful act as well. Now in general, we might think, if any permissible motivational state you can instantiate has to involve being most strongly motivated to do some particular thing, then it is rationally mandatory for you to be most strongly motivated to do that thing. In such a case, each way you can permissibly be is simply a partition of the different ways you could be in the rationally mandatory state of being most strongly motivated not to do the thing in question. We might thus refer to this as the:

> **Motivation Partition Principle:**
> If every rationally permissible motivational state that *A* can occupy involves being most strongly motivated to do *Y*, then it is rationally mandatory for *A* to be most strongly motivated to do *Y*.

So what we have seen, then, is that if you are in a rationally optional state of being sufficiently deterred from performing the wrongful action, then your current state and every other permissible motivational state you can be in involves being most strongly motivated not to perform the wrongful action. So it follows by the Motivation Partition Principle that if you are in a rationally optional state of sufficient deterrence from performing the wrongful action (say by overwhelming love of everyone), it is again rationally mandatory for you to be most strongly motivated not to perform the wrongful action.

Putting together our observations about the two ways in which you could be justifiably sufficiently deterred from performing the wrongful action, we can see that whether your state is rationally mandatory or rationally optional, it is still rationally mandatory for you to be most strongly motivated not to perform the wrongful action.

This covers the first (C21) case in which you are sufficiently deterred from performing the wrongful action and your state of overall motivation is justified. Let us turn, then, to the second way in which you could be sufficiently deterred from performing the wrongful action – that in which your state of overall motivation is unfitting or unjustified (C22). It might be, for instance, that the only reason you are currently motivated to avoid beating innocents is that you believe, falsely and irrationally, that Zeus will beat you if you do beat any innocents. If so, then it is rationally mandatory for you to get out of that state and to get into a different one. While your state of overall motivation is rationally impermissible, there will be some set of alternative motivational states, each of which would be rationally justified for you to be in (though this set may contain only one element – if it is rationally mandatory for you to be in one particular other state).

Niceties aside, for each rationally justified alternative to your current state, we will once again have two sub-case of what it will be like: (C221) the case in which the alternative motivational state is *not* a state of being sufficiently deterred from performing the wrongful act without your having to feel obligated not to perform it, and (C222) the case in which the alternative motivational state *is* a state of sufficient deterrence from performing the wrongful act in the absence of your feeling obligated not to. As an example of (C221), it might be permissible for you to be such that, after you shake your belief in Zeus, you are tempted to beat people from time to time. Now we have seen what happens with this sort of state. For given that it is morally wrong to do *X*, we have by our clarified analysis of wrongness and our Contour Thesis that this state of insufficient deterrence prior to feeling obligated can only be a permissible alternative if it also includes feeling obligated not to do *X* and being most strongly motivated not to do *X* as a result. Alternatively, as an example of (C222), it might be permissible for you to be such that, after you shake your belief in Zeus, you also happen to love everyone so much that you would never dream of beating them. But as we have observed before, such a state of sufficient deterrence to performing the wrongful action in the absence of feeling obligated not to do it is, *ex hypothesi*, a state of being most strongly motivated not to perform it.

Hence, we have seen that if you are in a rationally impermissible state of being sufficiently deterred from performing the wrongful action, your current state is

impermissible, but every permissible alternative motivational state that you can occupy involves being most strongly motivated not to perform the wrongful action. So we have by the Motivation Partition Principle that if you are in a rationally impermissible state of sufficient deterrence from performing the wrongful action (say out of unjustified fear of Zeus), it is once again rationally mandatory for you to be most strongly motivated not to perform the wrongful action. This covers the second (C22) case in which you are sufficiently deterred from performing the wrongful action and your state of overall motivation is unjustified. So putting together our observations about both the two general ways in which you could be sufficiently deterred from performing the wrongful action, we have that whether your state of sufficient deterrence is justified (C21) or whether it is not (C22), it is rationally mandatory for you to be most strongly motivated not to perform the wrongful action.

Finally, we can combine these observations about what follows if you are sufficiently deterred from performing the wrongful action with our previous observations about what follows if you are not to see that, if doing *X* is morally wrong, then whether you are already sufficiently deterred from performing *X* prior to feeling obligated not to do it (C2) or you are not (C1), it is rationally mandatory for you to be most strongly motivated not to do *X*. And from this, it follows by the Most-Motivation-Action Principle that you have conclusive reason not to do *X*. Hence we have that if doing *X* is morally wrong – which is to say that you are morally obligated not to do *X* – you have conclusive reason not to do *X*. Which is to say that the strong morality-reasons internalism thesis is true. (In the event that the reader would find it helpful to see the argument's steps more systematically, I have included this in the Appendix: Proof that there is Conclusive Reason Not to do what is Morally Wrong.)

It may be worth emphasizing what this kind of vindication of the strong morality-reasons internalism thesis does and does not show. On the account we have given, the fact that an act is morally wrong entails that there is conclusive reason not to perform it primarily because (1) for an act to be morally wrong, it must be the case that it is rationally mandatory to feel obligated not to perform it, and (2) in contrast to reasons for some other motivational states, if the reasons that count in favor of having to feel

obligated to do something are not sufficiently weighty to determine that we have conclusive reason to act out of them, they are also insufficiently weighty to determine that it is rationally mandatory to feel obligated to do it at all. Like other kinds of reasons for motivation and action, an agent's reasons to feel obligated to do things and to do them as a result of the fittingness of this motive can be overwhelmed by other considerations. It is simply that when these reasons are overwhelmed on the front of determining what it is rationally mandatory to be most strongly motivated to do, they are also overwhelmed on the front of determining what the agent must feel obligated not to do, and hence (given our analysis of moral wrongness) they are no longer sufficient to make it morally wrong not to act on them.

My vindication of the strong morality-reasons internalism thesis thus does not give any conceptual guarantee that any given consideration - even *I have promised* or *she will die if I don't help her* - is either a genuine obligation-making feature or a weightier reason than any other. It would show, however, that if we are as a matter of substantive normative fact morally obligated to do something, then we have conclusive reason to do it. Showing this, however, seems to be of some importance. For one thing, some have actually doubted whether an act's wrongness entails the existence of conclusive reason not to perform it. Sometimes, one might be quite convinced that something is definitely morally wrong, but wonder whether this really means that doing it is out, or that one has strictly most reason not to do it. I think that many of us are prone to do this in moments of weakness, or when it seems to us that morality demands of us some inconvenient or difficult thing. As Falk noted, the following can be a "perplexity of ordinary life":

> People commonly take it for granted that, when it is their duty to do some act, they have also a reason…for doing it, in some sense even an especially stringent one. But for many there comes a time, particularly when some personal interest seems at stake, when they feel troubled by doubt or in need of reassurance; and in this mood they turn to the moralist with a request: "Exhibit to me," they say in effect, "the reason…sufficient, even at cost to myself, to induce me to do what I ought, but don't want to do; for though I grant the duty, I see no such reason, and maybe there is none, or none sufficiently strong; and no one can do anything unless he has a sufficient reason for doing it, and knows just what it is!" (Falk 1948, 22).

What I have offered is a recipe for answering this question, and giving the requested reassurance. We need only ask the person who asks us this question for the

considerations that he takes to make the action in question "his duty," or something that it would be morally wrong for him to fail to do. Perhaps he has promised to do something, or perhaps someone will come to harm if he does not do it, or perhaps this thing he is thinking of taking or keeping is the property of another, or what have you. We then exhibit to the man our fitting attitude analysis of wrongness, and explain to him that he is committed to the view that the considerations he has given us makes it the case that it is rationally mandatory for him to feel obligated to perform the act in question if he is not already sufficiently motivated to do it, for otherwise, he would not have the duty he admits on account of the considerations he gives us. Next, we exhibit to the man our Contour Thesis and the Motivation Partition Principle, and explain to him how he is committed to the view that the considerations he has given us thus make it the case that it is rationally mandatory for him to be most strongly motivated to perform the act in question, for again, were it otherwise, he would not have the duty he admits on account of the reasons he has given. And finally, we exhibit to him the Most-Motivation-Action Principle, explaining to him, in the spirit of Chapter 3, that as the considerations he has given us count decisively in favor of being most strongly motivated to perform the act in question, they are *ipso facto* conclusive reasons to perform it.

If what we tell the man is true, I think that we will have succeeded in giving a maximally informative answer to his question. It succeeds in explaining to him how, if the act in question really is morally required, he has conclusive reason to perform it every bit as much as he might in certain circumstances have conclusive reason to avoid his own pain. Of course the man might waver in his conviction that the act really is, after all, morally required. But understanding the practical consequences of its moral requirement is no epistemic reason to waver so. Our evidence that we are morally obligated to do something is just as good as our evidence for any other normative fact, and it consists in the best unification of non-debunked normative intuitions by normative theories. There is no more reason to be skeptical of intuitions about wrongness in general than there is to be skeptical of intuitions about reasons for action in general, or to be skeptical in general of intuitions about epistemic reasons for belief.

I submit, then, that our explanation gives the man exactly the reassurance he asked for. It may not be the reassurance that he wanted. But even if philosophy can go

some small way towards replacing the hangman, we should be warned against thinking that it can entirely replace the therapist.[134]

4.6. Supererogation

Having thus explored the notion of what is morally good and the notion of what is morally obligatory or required, let us turn to the concept of supererogation, or action above and beyond the call of moral duty. One proposal for understanding supererogatory action would be to identify it with morally good action, or action that merits moral esteem. But it might well seem coherent to think that some actions are in no way morally estimable, yet supererogatory for all that. Suppose, for example, that I am in the position of the child in a slightly altered version of the case with the sniper from section 4.3. That is, I am drowning, and will soon die unless someone pulls me out. Over behind a wall is Bugsy, who knows that he will be shot dead if he emerges from cover and pulls me out. Let us suppose that Bugsy does this; he knowingly emerges, pulls me out, and is killed, thereby trading his life for mine. Let us suppose, however, that Bugsy did this solely because he is a member of a strange religious sect, and fervently believed that he would be sent to hell and tortured for all eternity if he ever let a stranger drown in his presence. His sole motivation was to do whatever he had to do in order to avoid this terrible fate – a

[134] Lewis (1996) rightly ridiculed the idea that we could explain why we ought to be ethical in terms of our states failing to deserve folk psychological names if we aren't. What is offered here to the consideration of the wicked is something quite different – namely that their view that something is morally wrong commits them to thinking that they have conclusive reason not to do - it in the very same action guiding sense as they might think they have conclusive reason not to do something on the grounds that it will avoid their own pain, secure their own pleasure, or what have you. This explanation would not be calculated to strike terror into their hearts, but rather to show them that, if they admit that doing something is morally wrong but go on to do it, they are either being logically inconsistent (thinking they do not have conclusive reason to avoid doing what their views analytically entail they have conclusive reason not to do), or that they are failing to do what by their own lights they have most reason to do (if they admit the conclusive reason against the act but perform it anyway). Since people are prone to avoid inconsistency and weakness of practical reasoning, coming to believe this can in principle play something of the deterrent role of the hangman. But of course people are often inconsistent, they are often practically weak or weak-willed, and they are often prone to ignore evidence if it suggests they must go against their inclinations. There is, I suspect, only so much a philosopher can do by way of pointing to the evidence and showing its entailments. There will likely be biases and defense mechanisms that one would need a sort of personal growth, of the kind people seek in therapy, to overcome. (Though of course there is no reason why acute philosophical attention cannot itself play this therapeutic role – it may even be particularly well suited to do so).

quick death followed by a nice afterlife beats torture for all eternity any day. Bugsy, we may suppose, did not care a fig for me on my own account; he feels strong obligations to his comrades but none at all to enemy civilians like me. Were it not for Bugsy's belief that he would definitely be tortured for all eternity were he not to help me, he definitely would not have done so.

In such a case, it seems very plausible (and certainly coherent) to think that Bugsy's action was in no way morally estimable. He was, after all, only trying to save his own irrational hide. But it also seems quite plausible (and coherent with the foregoing) to think that Bugsy still did more for me than what he owed me. He still, that is, did something more than what morality requires of him. If so, then we should want to say that Bugsy's action was supererogatory, but in no way morally estimable.

Another problem for the attempt to identify supererogatory action with morally estimable action is that it seems at least coherent to think that it could sometime be quite morally estimable to do just what morality requires. One example of this might be Gibbard's case of lashing out in grief, discussed in Chapter 2. If, for instance, you are extremely distraught with grief, you still might be morally required not to lash out at someone, but given how difficult it is, it might be quite morally estimable of you to contain your emotions and refrain from lashing out. Or, suppose you are carrying a wounded man to safety. It is not as though this is going to kill you – the cost to you is a rather arduous journey and a reasonable chunk of your time (a several hours to a few days). But let us suppose, however, that the man is constantly teasing you and harassing you; he feels entitled to your help, and he treats you as a total inferior. We might, in such a case, think that you are still morally required to help the man – it would, for instance, be morally wrong of you to leave him there to die merely on account of his treating you so badly. But while we might thus think that it was no more than what morality required of you, we might still think that your act of putting up with the man's harassment and saving him despite it was quite morally estimable all the same.

But if supererogatory action is not morally estimable action, then what is it? One natural suggestion in light of what we have seen might be to try to identify supererogatory action with action that it is rationally optional to feel obligated to perform.

This, after all, might seem to be the case of what might be the most plausible act of supererogation we have been discussing – namely that of Bugsy's getting shot in the head to save someone else from drowning. But this suggestion cannot be correct in its present form. For we saw that it might well be rationally optional to feel obligated to do things that do not go above and beyond the call of duty at all – if anything they were simply crazy, or well beneath its call. For we saw that it might be rationally optional to feel obligated to give all of drug *M* to the one man who needs the whole supply to live when each of the five others only need one-fifth. We also saw that it might be rationally optional to feel obligated not to put your mother in a home, when this might in fact be what you owe to your children and even to her (given that you cannot care for her adequately).

However, there is a natural way to amend the suggestion to cope with these cases. Supererogation is, after all, action above and beyond the call of duty. So whatever else it is, it must be action that is morally permissible but not morally required. A supererogatory act, then, is one which is such there is no rational mandate to feel obligated not to perform it, and such that there is no rational mandate to feel obligated to perform it. The requirement of moral permission might well be enough to keep things like giving the medicine to the one and not putting your mother in the home, which it may be rationally optional to feel obligated to do, from counting as supererogatory. So, one might try identifying supererogatory acts with those acts that are morally permissible, not morally required, and such that it is rationally optional to feel obligated to perform them.

But there is still a problem for this kind of analysis. For it certainly seems plausible (and at least coherent) to think that in the case where you are equidistant between the drowning swimmers, Bugsy and Suzy, it is rationally optional for you to feel obligated to save Bugsy, and it is rationally optional for you to feel obligated to save Suzy. But surely (or so it certainly seems coherent to think) both the act of saving Bugsy and the act of saving Suzy are each morally permissible and not morally required. Yet we might well not want to have to say on this basis that saving either of them is supererogatory – saving one or the other may be no more than what morality requires. One might think that something is going on here with the fact that there is more than one

option which it is rationally optional to feel obligated to perform which is morally permissible, and not morally required. But there are other cases of this description in which it seems plausible (and at least coherent) to think that both acts are supererogatory. For suppose we altered the case with the sniper so that Bugsy is equidistant between two different children drowning in two different ponds, where he can emerge from cover and pull one of the children out, but he will be shot dead by the sniper before he can pull out the other.

What we seem to need to understand an instance of supererogatory action is the idea that there was some act that the agent could have performed that was the bare moral minimum, and that the supererogatory act was strictly more favored by moral reasons than it. Following the understanding of an ethical reason for action developed in chapter 3, we might say that a moral reason for action is a consideration that contributes to the fittingness of a moral emotion and *ipso facto* counts in favor of having the motivations it involves and performing the actions that would realize the objects of these motivations. What makes an emotion a "moral emotion" is something we shall discuss a bit in the next chapter, but so far we have seen at least six instances of such: feelings of obligation, guilt, outrage, resentment, moral esteem, and moral disesteem. Now we should ask: are reasons to feel and act out of any of these emotions the kinds of considerations that make an act supererogatory or above and beyond the call of duty? I think that there is reason to doubt this. For one thing, it might be that if someone has done something culpable, it might be fitting to be moved by outrage or resentment to scold him or to refuse to help him in certain ways. Some such scoldings and denials of assistance might be morally permissible but not morally required. But surely, if they serve no further deterrent purpose or anything like that, we need not think that scolding the person or refusing to help him is any way supererogatory. These acts might be things that we are neither required nor forbidden by duty to do, but they do not go above the call of duty either.

I think that similar remarks may be made about reasons to do morally optional things (i.e. things that are neither wrong to do nor wrong not to do) out of guilt, moral disesteem, and moral esteem. Suppose, for instance, that Bill committed a crime some years ago. The crime was not too major (perhaps it was a property crime), but he cannot

find the parties against whom he perpetrated the crime to try to make recompense. But let us suppose that Bill is plagued by guilt for what he has done, and feels strongly inclined to confess to his crime and accept his punishment. As his past act was culpable, we might think that Bill is justified in feeling guilt for his act and being moved by it to confess his crime. But we also might think that enough time has passed and Bill has enough to live for such that it would be morally permissible to for him to let this go and get on with his life. We may thus think that Bill's confessing his crime and accepting his punishment is morally optional, and supported by the reasons that justify his feelings of guilt. But it would certainly seem consistent with our views for us to think that it would not be supererogatory for Bill to confess his crime. We think that confessing is neither required nor forbidden by duty's call, and we think that it is favored by a moral consideration (Bill's culpability), but we still might not think that it goes above and beyond the call of duty

Similarly, we might think that although it is rather disestimable to flaunt one's wealth in front of the poor, it isn't exactly morally wrong to fail to disguise one's wealth so as to spare their feelings. If so, then we have reason to be moved not to flaunt our wealth in front of the poor out of moral disesteem for the idea of flaunting it, although feelings of obligation needn't stand on guard to make sure that we don't flaunt it. But it might seem consistent with our views to think that refraining from flaunting our wealth isn't exactly supererogatory. Sure enough, we think that disguising our wealth is something we aren't morally required to do, but we might not think that it quite goes above or beyond the call of moral duty in the way a supererogatory act does.

Finally, suppose that someone breaks our lamp through no fault of his own, feels unfitting guilt for his action, and, out of such guilt, he apologizes profusely and offers to buy us a new one. The person is, we might think, doing something that is neither morally forbidden nor morally required, and in fact displaying a kind of punctiliousness that is rather estimable. If we are right about this, then there is reason to be moved to do as he does in a similar situation out of esteem for the idea of being so punctilious, though being that punctilious is not something we need to feel obligated to do if we are tempted not to do it. But it might seem to be consistent with our views to think that that apologizing and offering to replace the lamp isn't exactly supererogatory either. It is not morally

required, and it is nice, but it may not quite seem to go above and beyond the call in the way supererogation does.

It seems, then, that the kinds of moral reasons that can make one of several morally optional acts morally permissible are not just any considerations that count in favor of having and acting out of a moral emotion. In our example with Bugsy and the sniper, we saw that considerations that make feeling obligated to do something rationally optional (but not rationally mandatory) seem to make for supererogatory action. But we also saw that considerations that contribute to the fittingness of outrage, resentment, guilt, moral esteem, and moral disesteem do not seem to fit the bill. It would appear, then, that we might identify those moral considerations that make an act supererogatory with those that contribute to justifying, but not mandating, feelings of obligation to perform the act. As such, the foregoing observations would support the following analysis of supererogatory action in terms of the fittingness of feelings of obligation:

> **Fitting Attitude Analysis of Supererogation**:
> To judge that it is supererogatory for agent *A* to do *X* is to judge that:
> (1) It is morally optional for *A* to do *X*
> (That is, it is morally permissible but not morally required for *A* to do *X*. In terms of fitting attitudes, this means that it is neither rationally mandatory for *A* to feel obligated *not* to do *X* if *A* insufficiently motivated to *not* do *X*, nor rationally mandatory for *A* to feel obligated *to* do *X* if *A* is insufficiently motivated *to* do *X*)
> (2) It is rationally optional for *A* to feel obligated to do *X*, and
> (3) There is another morally optional act, *Y*, such that *A* has strictly more reason to feel obligated to do *X* than *A* has reason to feel obligated to do *Y*.

Thus, if an act is supererogatory, it is fitting (but not mandatory) to feel obligated to perform it. Since, as we have seen, feeling obligated to do something involves motivation to do it, it follows by the Warrant Composition and Motivation Action Principles that an act's status as supererogatory entails that one has reason to perform it. Moreover, since the reasons to feel obligated to perform the supererogatory action are

strictly stronger than the reasons to feel obligated to do other things, one has strictly more reason of this kind to perform the supererogatory action than to perform the other action. For instance, suppose that Bugsy's two morally permissible options are to either sit tight and survive or save the drowning child at the cost of his life. If the latter is supererogatory, he has strictly more moral reason, or reason of the kind that can make acts morally required or supererogatory, to save the child at the cost of his life than he does to sit tight.

Supererogatory action is thus action that is supported even more strongly than merely permissible action by the same kinds of reasons that count in favor of doing what is permissible. By understanding supererogatory actions that those that we have even stronger reason to perform out of feelings of obligation than we have to perform actions that are merely morally required, we uncover a clear sense in which supererogatory action is action *above* the call of duty. As such, we can actually understand supererogatory action in a gradational and not merely a binary way: an act is more or less supererogatory to the extent that it is supererogatory, and more or less strongly favored by reasons that make it optional to feel obligated to perform it. Thus, someone might coherently think that sacrificing your legs to save Bugsy is supererogatory, but that sacrificing your legs and your life to save both Bugsy and Suzy is even more supererogatory. The idea here would be that the fact that *Here is Bugsy and he needs the help I can provide* counts in favor of feeling obligated to save Bugsy, and the fact that *Here is Suzy and she needs the help I can provide* counts in favor of feeling obligated to save Suzy. But together these reasons favor the act of saving both Bugsy and Suzy (and getting killed) more than they favor the act of saving only Bugsy (and surviving but losing your legs).

But while our reasons to feel obligated to do supererogatory things support doing them even more strongly than they support our doing things that are merely permissible, there is no guarantee that we will have more reason overall to do what is supererogatory (or most supererogatory). For although there is more moral reason to perform supererogatory actions, there may well be more non-moral reason not to. Obviously: the fact that something will get you killed (when you have a life worth living ahead of you) is a strong reason not to do it, and it is certainly coherent to think that this reason is at least

as weighty, if not weightier, than your reasons to save the child on account of her needing your help. We have seen that rational mandates for feeling obligated to do something entail the existence of conclusive reason to do it, but along the way we saw that rational options for feeling obligated to do something do no such thing. It is certainly coherent to think, then, that there are cases in which you have most reason not to do what is supererogatory, even though it is what you have most moral reason to do.

We have seen, then, that there is the following structure to the relationship between morality and reasons for action. We have the notion of cases in which our reasons to feel obligated to do something succeed in making it rationally mandatory for us to feel this way, and by doing so guarantee that we have conclusive reason to do it. We also have the notion different kinds of cases, in which our reasons to feel obligated to do something succeed only in making the feeling of obligation optional, and guarantee only that we have some reason to do it. We could say, then, that the first way in which moral considerations make a claim on our actions is that of a demand that we rationally must meet. In the same spirit, we could say that the second way in which moral considerations make a claim on our actions is that of a recommendation that we may comply with but (if there are indeed at least equally strong reasons to act to the contrary) need not. Obligatory or morally required acts are those that we have moral reason to do, and most reason to do, but not necessarily most moral reason to do. The most supererogatory or most highly morally recommended acts, on the other hand, are those that we have moral reason to do, and most moral reason to do, but not necessarily most reason to do. So long as some acts are obligatory and others are supererogatory, morality goes only so far with demands, then tries to go even further, but does so only by way of recommendations.

4.7. Moral Blameworthiness and Retributivism

Let us conclude by turning from some of the nicest things that moral considerations might support to some of the meanest. The framework we have been developing so far is

one of understanding moral concepts in terms of the fittingness of moral emotions, and understanding morality's relation to reasons for action by means of the connection between the fittingness of motives and our reasons to realize their objects. This framework is, I think, quite germane to articulating the guiding thoughts behind both retributivism and several other kinds of thoughts about our reasons to discount the interests of the culpable, which I think are often misunderstood. With it, we can see how the justification of our moral emotions of outrage, resentment, and guilt might entail the existence of reasons to do and allow some things that are not so very nice. There are, however, reasons to doubt whether these "reactive attitudes," the fittingness of which would constitute reasons to do such things, are ever really justified. Luckily, our understanding of reasons to comply with the demands and recommendations of morality shows *these* reasons to be quite unthreatened by the possible unfittingness of guilt and anger. Once we sufficiently appreciate the distinction distinguish between moral reasons and moral responsibility, we can see that the non-existence of the latter might not be so catastrophic.

Start first with the idea of retributivism, that the mere fact that someone has done culpable wrong, and consequently "deserves punishment" can give us normative reason to punish her.[135] There is no reason why this view must be confined to discussions of official punishment by state institutions; we could apply such a view to situations of groups of people stranded on desert islands, or to everyday personal interactions. "Punishment," then will be something that goes against the interests or autonomy of the person who is punished, like flogging or house arrest, but it will also include much more mundane things that we can do to each other (often quite within the confines of the law) like scolding, kicking out of the house, withholding desert, and so on.

Retributivist intuitions are probably most easy to get in both the gravest cases and the least grave cases. For instance Moore (2000) describes several instances of absolutely heinous instances of torture and murder, sometimes with an offender who shows little or no remorse, and says so. In such cases, some of us can get the intuition

[135] See for instance (Moore 2000, 747). Often, the view is put so that it includes as a separate clause the view that culpability is a necessary condition for permissible punishment. But as we shall see, some the view's main explanatory work might be accomplished by means of the clause about there being positive reason to punish the guilty alone.

that there is reason to "make this person pay," whether or not anything else comes of it. Whether or not, for instance, it deters anyone, teaches anyone anything, provides entertainment or solace, or what have you. On the other end of the extreme, we might consider cases in which your roommate has ruined something of yours in the course of playing a stupid joke. It can seem that this justifies yelling at him, refusing to do him the favor you were going to do him, or even playing a nasty joke on him or messing up one of projects just because of what he did to you. Quite independent of whether any of this will actually do anything further, like improve his future treatment of you or anyone else, there can seem to be reason to do these things that he finds aversive just to "get even."

One thing that can seem to lend credibility to retributivist thinking is its potential to explain some other things that we find plausible that might be otherwise difficult to account for. As has been well remarked since Carritt (1947), we tend to think that it is permissible to punish the culpable in order to achieve social aims that it would be impermissible to achieve by punishing the innocent. Thus we might think that the fact that imprisoning Smith will deter crime to degree *X* is a sufficient reason to imprison him if Smith is guilty of a crime, but not if he is innocent of the crime. And this is so taking into account all such factors as a guilty Smith's possibly being more likely to offend than an innocent Smith. Or again, consider how severely we tend to think we should punish someone. As Pogge (1995) has pointed out, we might be able to greatly increase the net number of people who survive if we were to execute a certain number of people caught drunk driving. But the savings in lives seems to us an insufficient reason to punish them so severely. Since we tend to think that drunk drivers have done something culpable and deserve some punishment, we tend to think that considerations of social utility like deterrence can count decisively in favor of giving them the punishments they deserve. But because we do not feel that they have done anything so culpable as to deserve execution, we do not think that considerations of social utility can justify doing this thing to them that they do not deserve.

Or consider even cases of discounting the interests of the culpable that are quite outside of what we would typically regard as punishment, like harming in self-(or other-) defense. Many of us seem inclined to think that we are allowed to do things to the culpable in the course of defending ourselves that we would not be allowed to do to the

innocent. If, for instance, someone is culpably trying to kill you, we tend to think that you are quite allowed to kill him so as to prevent him from doing so. But for many of us this does not appear to be because we think that if it comes to a choice between your life and someone else's, you are allowed to do whatever it takes to save your own. For as Boorse and Sorensen (1988) and McMahan (2002) point out, many of us do not think that it would be permissible to do certain things that get innocent people killed in order to save our own hide – for instance by grabbing an innocent bystander and using her body to absorb the bullets that are intended for you.

By positing a general sort of positive reason to act against the interests of the culpable on account of what they have done, retributivism might seek to account not only for our direct intuitions in favor of punishing them in the absence of any further reasons to do so, but also our intuitions about the extent to which social utility can justify punishing the guilty but not the innocent, as well as those about our permissions to do things to the guilty but not the innocent in the course of defending ourselves. If there is already reason to imprison a guilty man, then when considerations of social utility come on the scene, they can combine with the former to make the imprisonment justified all things considered. But if the preexisting reasons to imprison the man are absent, as they are in the case of the innocent, the considerations of social utility may not be able to tip the balance in favor of imprisonment. Of course, the retributivist might seem to do much better if he could somehow make out why culpability is both a positive reason to punish someone and (which might be more intuitively plausible) a consideration that weakens the strength of existing reasons not to harm the culpable above and beyond its counting positively in favor of punishing them. For we do not usually think mere attempted murder to be something that warrants killing someone (even, we might think, if it could save others, say by means of deterrence). It would appear that the retributivist would do better in explaining the permissibility of killing in self-defense if he could say that the culpability of the attacker lowers the barriers to harming him in a way that is disproportionate to its strengthening one's positive reasons to harm him.

Such, then, are some of the intuitions in favor of discounting the interests of the culpable, and the ways a retributivist might try to account for them. But retributivism, as

well as the intuitions about reason to punish the culpable simply on account of their culpability that most directly support it, can seem puzzling. "What is the point" some seem to wonder, "of harming someone just because 'she deserves it' or 'she has it coming' or 'to get even with her?'" Of course, the point of such acts of harm, according to the retributivist, is just that inflicting it would even the score, or that the guilty party deserves it, or that the guilty party has it coming, or what have you. But what the "what's the point" question really communicates is a question about why the things that the retributivist claims to be worth pursuing as ends would be worth pursuing as ends.

Along these lines, retributivists have attempted to point to our feelings of outrage or resentment and guilt,[136] but I think that it has often been unclear what about these attitudes might help explain how meting out deserved punishments could be an end in itself. As often as not it may be thought that reference to feelings of anger can only mean a view according to which punishment is thought to be justified by its bringing about satisfaction to angered parties by allowing them to vent their rage. But, of course, this is not the retributivist idea, for then it would not be the bare fact that punishment was deserved that gave us reason to punish, but the fact that the punishment would satisfy someone's desires. Instead, we should look to the understanding afforded by the fitting attitude analysis of blameworthiness and the connections we have developed between fitting attitudes and reasons for action.

As we have remarked, feelings of outrage and resentment, as species of anger, involve motivations to behave aggressively or punitively towards those who are their objects. There is a good range of behavior that can count as aggressive or punitive, but in general it involves harsh treatment, ranging from physical harms, to denials of assistance, to bossing someone around, to scolding. Now what is very important to note is that the punitive motivations anger involves are genuinely intrinsic, or non-instrumental. When you feel angry at someone and you feel like punching him, you are motivated to do this as an end in itself. Of course, you can simultaneously be motivated to do it in order to achieve something else, but the motivation to punch involved in feeling angry is not a motivation to punch in order to bring something else about – other than, perhaps, the

[136] For a discussion of anger in this context see for instance Murphy and Hampton (1988), and for a discussion of guilt in this context see Moore (2000).

person's feeling pain, or feeling humiliated, or in some other way being aversely affected. Similar remarks hold for motivations to deny assistance, restrict autonomy, and scold: in being motivated to do these things out of anger, you are motivated to do them on their own account, or simply on account of the harm they do to the object of your anger. This may not be pretty, but it is, I submit, what it is to be motivated by anger.

Now, by the fitting attitude analysis of blameworthiness, we have as an analytic truth that if a person has done something morally culpable, it is fitting for others to be angry at her. In more detail, we might say, it is fitting for those who have been wronged in particular to feel resentment, and fitting for those "feeling on their behalves" to feel this as well, while it is fitting for others to feel outrage. By way of example of what it is to feel on someone's behalf, we would probably tend to have a species of anger that somehow felt "a bit more personal" towards someone who had wronged our infant our companion animal than we would if we heard of someone's wronging a stranger's infant. Of course, we can probably feel resentment on anyone else's behalf too, if we just do enough sympathizing with them or imaginative simulation of what it would be like to be in her situation. These distinctions between resentment and outrage are no doubt interesting, but for my purposes they will not matter much; I mention them only so as to explain what is meant more precisely when I speak of a blameworthy act befitting "outrage and resentment."

Now since resentment and outrage involve intrinsic motivations to behave punitively, we have by the Warrant Composition Principle that if it really is fitting to feel resentment or outrage towards someone, then it is fitting to be intrinsically motivated to behave punitively towards her. But we also have by our Rational Ends Principle that if it really is fitting for us to be intrinsically motivated to behave punitively towards someone, then we have reason to do so an end in itself. The lesson of Chapter 3 was that what it is ultimately worth pursuing as an end in itself is just what it is fitting to be intrinsically motivated to pursue. This, I argued, best explained the structure of practical reasoning in general and its guidance of motivation and action. The things we have discussed so far as examples of things that it is fitting to be intrinsically motivated to do and thus worth doing as ends in themselves have been rather nice, or at least neutral – avoiding your own pain, preventing the pains of others, saving people's lives, not breaking promises, and so

on. But our general principles will allow no exception: if it really is fitting to be intrinsically motivated to harm some person, we really must admit that it is worth aiming to do it as an end in itself.

Of course, even if it is worth harming the culpable as an end in itself, our reasons to do so might be ever so much weaker than our reasons to do other things like help people and prevent the innocent from being harmed. But this is something a retributivist might be happy to admit too, as with "better to let a hundred guilty men go free than that one innocent should suffer." It may even be that our reasons *not* to harm the culpable as an end in itself are stronger than our reasons *to* harm them as an end. And surely this is something a retributivist might want to incorporate in making sense why we should often show mercy to the culpable. But it suffices to make sense of the basic retributivist idea, and to respond to the basic puzzle about it, to see that culpability entails the fittingness of intrinsic motivation to punish, and the fittingness of intrinsic motivation to punish entails intrinsic reason to do so, or reason to do so as an end in itself.

In conjunction with the features of moral obligation that we noted earlier, the retributivist might try to use the foregoing vindication of his view to capture our intuitions about greater moral latitude we have in inflicting harms on the culpable in the pursuit of socially desirable ends. For if we have these intrinsic reasons to harm the culpable, it will be more likely that the considerations in favor of having to be most strongly motivated not to harm them will be indecisive, and thus that it is not rationally mandatory to feel obligated not to harm them. In conjunction with the Contour Thesis and our analysis of wrongness in terms of rational mandates for feelings of obligation, this will entail that it will be less likely to be morally wrong to harm the guilty in bringing about some social good.

Similar considerations might be used to explain why the culpable themselves have less moral latitude in resisting the punishments we mete out to them than the innocent would. For the fitting attitude analysis entails not only that it is fitting for those other than the offending party to feel outrage or resentment – it also entails that it is fitting for the offending party himself to feel guilt for what he has done. Now as we saw guilt involves motivation to make amends, but it also involves motivation to accept

punishments, or to allow others to punish you without your tending to get angry yourself. Indeed, we might observe, feeling guilt involves *intrinsic* motivation to self-punish or to accept the punishments of others. When you feel guilt, you are not simply moved to accept your punishment because you think it will have nice consequences (like the restoration of social bonds and the fact that both you the punisher will feel better in the long run); you actually feel moved to accept your punishment as an end in itself. As such, our Warrant Composition and Rational Ends Principles will entail that if you really are culpable for doing something, then you really do have reason to let others punish you for it as an end in itself.

Now both the guilty and innocent alike might have strong reasons to avoid harming non-culpable others (like police with good evidence that you are culpable) in the course of trying to resist punishments. But if you are innocent, we might think that sometimes your reasons to resist the punishment justifies inflicting such harm. We might think, for example, that an innocent person with no way to prove his innocence would be permitted to knock a guard over the head in an attempt to escape from prison. If the guard learned that the man was innocent, he might well be inclined to think that he had no right to resent the injury.[137] On the other hand, we might tend to think that the guilty would not be morally permitted to do this. If a guard were knocked over the head by a heinous criminal serving a sentence he richly deserves, he might well think that he had reason to resent the injury. We have on hand a way to make sense of these views. For if you are innocent, you will have no intrinsic reason to allow others to punish you. As such, it will be more likely that the considerations in favor of having to be most strongly motivated not to harm others in the course of resisting punishment will be indecisive, and thus that it is not rationally mandatory to feel obligated not to do so. In conjunction with the Contour Thesis and our analysis of wrongness, this will entail that it will be less likely to be morally wrong for you to harm the others in resisting punishment if you are innocent than it will be if you are culpable.

[137] Similar attitudes might be taken towards attempts by captured enemy combatants to escape in time of war. The idea again might be that (especially if they are draftees, or otherwise not so culpably involved in the persecution of an unjust war) that as prisoners who do not deserve their imprisonment on account of some culpable wrong, they have more moral latitude to resist it.

Is there, however, a way in which a retributivist might draw on our analysis to explain not only how there are positive reasons to harm the culpable, but how culpability can give rise to what Dancy (2004) calls "weakeners" of reasons not to harm them, or "strengtheners" of reasons to harm them? A weakener of a reason not to harm the culpable would be a consideration that decreases the strength of our reasons not to harm them without constituting an additional positive reason to harm them. As we saw, weakeners of reasons not to harm the culpable might seem to be better suited than positive reasons to harm them for purposes of explaining the apparent permissibility of killing the culpable but not the innocent in self-defense. One way in which a retributivist might try to avail himself of this sort of explanation would be to say that being angry at someone involves not only motivation to punish or harm her, but also a tendency to be less inclined not to harm her. Perhaps part of what it is to be angry at someone is to tend to be less motivated by feelings of care or obligation or what have you not to harm her. If this is right, then our retributivist might be able to make out, by means of the Warrant Composition Principle that it is fitting to tend to be less moved not to harm the culpable. From here it would follow that the overall motivational states it is fitting to be in given the fittingness of your anger will tend to involve less motivation not to harm the culpable, which is exactly the kind of weakening we are looking for.

We have seen, then, that the fittingness of outrage, resentment, and guilt, will vindicate the basic retributivist idea that punishing the culpable is worth pursuing for its own sake, and will help retributivism to make sense of the idea that we are allowed to do things to the culpable for further ends that we are not allowed to do to the innocent. But are these attitudes ever really fitting? A claim so bold as one that an entire class of attitudes is never actually warranted would seem to carry with it a strong burden of proof. But there are reasons to suspect that this burden could be discharged.

Start with our ordinary thinking about the difference between wrongness and culpability. As we saw with Gibbard's case of lashing out in grief, we often excuse wrongdoing when the wrongdoer was less than fully responsible. Importantly, we don't take diminished responsibility to justify what would otherwise be wrong. It is not as though, given that one's self-control is diminished, one somehow has less reason not to

lash out than before. It is simply that since we know that you are not fully in control of your failures to do what you have most reason to do, we do not blame you for them. We do not think that you deserve the same harsh treatment that you might if you were fully responsible (like confrontation to the effect of "hey, don't get mad at me, bub, I was only trying to help!").

It is thus important to distinguish thoughts about exculpation from thoughts about the non-existence of reasons to do impossible things. Reasons for action apply to what is in your voluntary control. Even if altering the planetary orbits in a particular way would achieve all of your rational aims better than anything else you could do, you still have most reason to carrying on doing things other than altering the planetary orbits, as this is something that is outside of your voluntary control. Our reasoning and planning applies only to what it can direct in principle; if you know that your reasoning and planning cannot even in principle affect the orbits, you cannot consistently reason or plan to do so.

Similarly, if a mad scientist got control of your peripheral nervous system and started using you like an unwilling marionette to do horrible things, you would have reason to try to regain control of your peripheral nervous system, but you would not have reason at the time of action to omit doing what is done by the bits of your body that remain outside of your voluntary control. For avoiding doing these things would be to you then like altering the planetary orbits is to you now. As the mad scientist uses your right hand to strangle someone, you have most reason to do whatever you can to prevent the hand from doing this, but you do not have most reason to simply let go.

Loss of actual voluntary control of something thus changes what you have reason to do, and with this it changes what it is morally wrong for you to fail to do. Indeed, it is most natural to say that when my hand is directed by the mad scientist rather than my intentions, what it does is not even my action, and what it does thus cannot count as something it is morally wrong for *me* to do. But this is not how we typically think about mere exculpation. Exculpation pertains to failures to do things that were still in principle within the control or scope of causal influence of our intentions and other normatively governed mental states. When a typical person lashes out in grief, it is not as though all possible causal links are severed between her reasoning, her motives, and her action.

Lashing out is thus something that she could sensibly have judged herself to have reason not to do, felt obligated not to do, been motivated not to do, and so on.

Thus, our thinking about what it is morally wrong to do, or what we should feel obligated not to do, presupposes only what is presupposed by all of our thought about what to aim at and what to do, namely that some of what we aim at and do is within the scope of influence of our thoughts and intentions. But thought about what is culpable to do, or what we should feel guilty for doing ourselves and feel outraged or resentful at others for doing, presupposes something stronger. It presupposes that we are responsible for a failure to do what we have reason to do in some sense stronger than simply that the action we failed to perform was within the scope of influence of our intentions. One way it seems that you can fail to be responsible in this stronger sense is for your motivations to be recalcitrant, and for your rationally guided intentions to be too weak to direct your conduct in the face of these recalcitrant emotions. Thus we might not blame the person deep in grief for lashing out because we take it that his anger was recalcitrant to his judgments, and this recalcitrant anger was too strong for his intentions to determine what he did. But of course some such explanation might seem to be available every time someone fails to do what he thinks he should do. For every time you fail to do what you thought you should, your views about what to do were, *ex hypothesi*, been overpowered by some other causal force. Does this mean that every time you fail to do what you think you should your intentions are overwhelmed in such a way as to exculpate you?

One might not think so. One might agree that *sometimes* one's recalcitrant emotions are so strong that "no matter how hard you try" to do what you think you should, the causal influence of your views about what to do will "almost surely" be overwhelmed by your recalcitrant attitudes. But, the contention would be, on other occasions, "if you just put forward enough effort," the causal influence of your views about what to do can overcome that of your recalcitrant emotions. This is, I take it, exactly the common sense view about when weakness is exculpated and when it is not. But we need to ask: what exactly determines whether or not you put forward the requisite effort? For if you are not in control of how much effort you put forward, if how much effort you put forward is just something that happens to you without your having any say

in the matter, then we might well think that it would not be fair to be angry at you for what you do as a result of failing to put forward the requisite effort.

Remember that in asking questions of responsibility and exculpation we are not asking about pragmatic reasons for anger. We are not asking about whether or not it would bring about some good state of affairs if we were to be angry at you. We are asking whether you are blameworthy, which requires that it be *fitting* for us to be angry at you. It requires that you deserve our anger, and in fact, given the fact that anger involves intrinsic motivation to punish, and the Warrant Composition and Rational Ends Principles, this is a question of whether we have reason to punish you for what you have done as an end in itself. So this question of whether it is fair to be angry at someone seems not to admit of such a flippant response as "of course it's fair because angering in this way brings about these goodies." For whether or not it is good policy to get angry does not speak to anger's fittingness and the question about rational ends that goes along with it. You cannot justify harming someone as an end in itself by pointing to the benefits of doing such a thing.

With this in mind, I think that we should be more ready to admit that it would be unfair for us to be angry at you for doing something unless you were somehow in control of the degree of effort that you put forward in trying to do it. For it might seem quite barbarically false to think that we have reason to harm you as an end in itself for doing something determined by a degree of effort over which you have no control. What, we might wonder, would be the normative difference between harming someone as an end for something beyond her control and beating someone's innocent relative simply in order to get back at him for something that he did?

There might be two main candidates for how you control the degree of effort *E* that you put forward in trying to do *X*. First, putting forward *E* might itself be a voluntary action, which you control by putting forward some second degree of effort, *E*'. But if it is unfair to be angry at you for doing *X* unless *E* is in your control, how could it be fair to be angry at you for putting forward *E* unless *E*' is within your control? And if it takes an *E*' within your control in order for you to be responsible for *E*, then surely we will need an *E*'' – a degree of effort you put forward in putting forward *E*', over which you had control by means of an *E*''', and so on *ad infinitum*. It might be a

metaphysically interesting question whether there could be a being with an infinite hierarchy of actions of putting forward effort, but it seems pretty clear that we are not that being. So if we are to be in control of the degrees of effort we put forward in performing an action, they must be controlled by something other than further degrees of effort.

A second main way in which we could control the degrees of effort we put forward is by means of our views about whether we should put them forward. The degree of effort we put forward to do something might be a motivation of a certain strength, which we govern by means of our judgments about whether or not that degree of motivation is fitting. But again, if it is fair to be angry at us for what we do as a result of our normative beliefs about how we should be motivated and what we should do, it seems that we had better be responsible for those normative beliefs. Indeed, we intuitively recognize here a distinct way in which a person can be exculpated in addition to her being overcome by "uncontrollable" passions. For we often think that a person can be excused for doing something that he should not have done on account of the fact that he "did not know any better" – that is, on account of the fact that he failed to have correct views about what he had reason to do in such a way that this failure to know what to do was itself was something for which he was not (fully) responsible. Thus, we tend to think that people who are brainwashed are less culpable for their wrongdoing than those who are not so coercively indoctrinated, or that those who have not had certain educational experiences that would enable them to have the right views are less culpable for their wrongdoing than those who have.

But what exactly is it to be responsible for your views about what to do? A natural thought might be that you are responsible for them to the extent that we can trace them to past decisions about what kinds of inquiry to undertake, or decisions about when to reflect and think things through (and perhaps how best to conduct such inquiry). But what if your acts of thinking things through were themselves out of your control? If they were, can we really say that you are responsible for the views you hold now as a result of whether or not you did so? Recall here the stakes associated with the notion of responsibility in question. What we are really asking is whether it is fair to blame you for coming to the wrong views if what led you to them were instances of thinking things through which were themselves out of your control, or for which it would not be fair to

blame you. We might well demur, and admit that for you to be responsible in the relevant sense for your views about what to do, it must indeed be the case that you were responsible for the acts of thinking things through that led you to them.

But now it seems the game is up. For, as with any other kind of act, we will have to ask what makes it the case that we are responsible for an act of thinking things through. The answer we have come to entails that we are responsible for acts of thinking things through to the extent that we are responsible for the beliefs about whether we should think things through that led us to do so or not, and that we are responsible for these beliefs to the extent that they themselves trace to past acts of thinking things through that led us to them. It might be a metaphysically interesting question as to whether these conditions for responsibility in action could be satisfied by a being who had been deliberating about how to deliberate for all eternity. But the obvious and practically important fact of the matter is that we do not satisfy these conditions. There was something about how our brains were constructed, or how we were acculturated, or some set of quantum events, or something of the sort, over which we most definitely did not have control, which put in place our first thoughts about whether or not to think things through. So trying to ground an agent's responsibility for her actions on the foundation of her responsibility for her past actions of thinking not only leads to a potentially vicious regress – it actually seems guaranteed to give us the result that we are not responsible for our actions.

We have thus seen reasons to think that in order for someone to be responsible for her action in a way that would justify anger at her if it is wrongful, she would have to be in control of the mental states that lead her to the action, but she would also have to be in control of those, and so on, until we reach states over which she definitely has no control. For my part I can see only two ways to resist the conclusion that people are not responsible for their actions in a way that makes it unfitting to be angry with their wrongdoing. The first is to embrace Barbarism, or the view that it is fitting to be angry at – and to punish people as an end – for things that are out of their causal control. The second would be to insist that at some point an agent's control of her behavior and mental states is not a matter of causation by others of her mental states at all, but causation by "her" more directly. This would be something like Libertarian free will, but for our

purposes it will only make sense to distinguish Libertarianism if it is more than Barbarism conjoined with metaphysical profligacy. So it will not be interesting to simply say that the way behavior gets generated is by means of this or that kind of brute fact about what people think they should do or intend to do, or that it gets caused by these or those un-law-like activities of their soul-pellets. To be different from Barbarism there must be something about this metaphysical picture that enables us to make genuine sense of an agent, rather than certain brute facts, being in control of her behavior in such a way that it is fitting to be angry at and punish her as an ultimate end if her behavior is wrongful.

But I definitely do not intend to settle this. I merely walked through the foregoing to try to make plain how there are what I take to be very good reasons to suspect that our attitudes of outrage, resentment, and guilt must always and forever be unfitting. In fact, if the foregoing is correct, it may actually be that, *pace* Strawson (1968), an understanding of moral responsibility in terms of the potential fittingness of emotions like outrage and resentment contributes to undermining, rather than vindicating, the idea that we are morally responsible. For so long as we ask questions of moral responsibility in the abstract, it is comfortable enough to try to draw a line somewhere in the web of mental causes of behavior and say "yes, I see that these states of yours are determined by events, laws, and chances beyond your control, but we shall still regard whatever these states produce as your doing, and we shall hold you morally responsible for it."

Indeed, if we are not careful about what is at stake, we might even be inclined to be so flippant with questions of moral responsibility as Strawson himself, who first recognized a distinction between the fittingness of a reactive attitude and pragmatic reasons to have it, and then offered blatantly pragmatic reasons to continue to be angry at people. For Strawson embraced the idea that the following was a "lacuna" in the "optimistic story" of moral responsibility:

> The only reason you [the optimist] have given for the practices of moral condemnation and punishment in cases where this freedom [i.e. freedom in the sense of "the identification of the will with the act," or acting on one's motives and normative views] is present is the efficacy of these practices in regulating behaviour in socially desirable ways. But this is not a sufficient basis, it is not even the right *sort* of basis, for these practices as we understand them (Strawson 1968, 74).

After this Strawson spends much time arguing that the attitudes which he dubs "reactive" are part of our normal interactions with each other. It is quite unclear to what extent some of his examples really raise the same kinds of questions as those raised by outrage and guilt. To have things pretty much like "love" as we know it between two psychologically typical adult humans, we might not require anything more than esteem. Yet there are reasons to suspect that it may be fair to esteem and disesteem people for how they are even if how they are is outside of their control in a sense more demanding than simply that of "within the scope of influence of their normative views." Also, anyone who is close to young children, or to someone who is mentally disabled or mentally ill, can probably appreciate that some of our most important social relations do not require views to the effect that both parties are apt targets of anger. If one then imagines this element of those relationships in combination with various of the most central elements of adult relationships (like shared interests and activities, joint inquiry or reasoning, and so on) one might be forgiven for being rather puzzled by the claim that we need to think emotions like guilt and anger fitting in order to make our social relationships at all recognizable. But be this as it may, Strawson does not content himself with saying that emotions like anger are so central to our lives that we are simply stuck with them. He goes on to say:

> If we could imagine what we cannot have, viz. a choice in this matter [as to whether to we should continue to have attitudes like anger], then we could choose rationally only in light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of this choice (Strawson 1968, 84).

The uncharitable reading of Strawson is that he differs here from the "optimist" he criticized above only in that he uses more ethereal sounding language to describe the goodies anger is supposed to get us: the optimist spoke of 'the efficacy of practices in

regulating behaviour in socially desirable ways,' while Strawson speaks wispily of 'the gains and losses to human life, its enrichment or impoverishment.' The charitable reading of Strawson is that his optimist spoke of reasons for attitudes whereas Strawson is speaking of reasons to make ourselves such as to think that we have reasons to have attitudes. But whichever way you slice it, if I am right, there is something here that the issue Strawson calls "the general thesis of determinism" bears upon. This is that, if people take their anger to be fitting (and they realize a close-in entailment of their view) they will take themselves to have intrinsic reason to punish its object. And this sort of thing tends to get people intrinsically punished. This may be a fuddy-duddy way of looking at things, but I would think that the question of whether we should let ourselves continue to think that anger is fitting has a lot to do with whether such thoughts are true, and thus whether we really do have intrinsic reason to do the kind of punishing that we take ourselves to have in thinking anger fitting. But this question of whether anger really is fitting and whether people really deserve punishment cannot be answered by questions about social goodies, no matter how airily or earthily you describe them. You need to ask whether it is really fair to anger at and aim at harming people on the grounds of what they have done, and I see no way to answer that without asking questions about the kinds of control people can and cannot have over themselves, whether as a matter of nomic, metaphysical, or conceptual necessity.

Thus, once we understand the connections between the fittingness of the reactive attitudes of outrage and resentment and that of intrinsic motivations to punish, and the connection between this and intrinsic reasons to punish, we appreciate the awesome gravity of attributions of moral responsibility. And when we appreciate this, we are less apt to be flippant about the question of where to draw a line in the network of mental causes. We could of course reform our concept of moral responsibility if we wanted, but we should be very careful in communicating the results. For if I am right, we currently have a concept that is as practically important as it may be pernicious – it is that which ties in with reasons to harm people as an end in itself. If nothing really falls under that concept, it may be very important to explain to people that that this is so, and to make sure that they do not take things cast in the language of the reform to justify harming people in ways they do not deserve.

So we have seen how the idea behind retributivisim would follow from the fittingness of feelings of outrage, resentment, and guilt, and we have seen how we have reasons to doubt that outrage and resentment (to which we could just as easily add guilt, for the same reasons) are ever fitting. We might say that the price of freeing ourselves from the retributivist idea is thus error theory about moral blameworthiness. But how much of a price is this to pay? There may well be all kinds of important implications for our substantive ethical theorizing about what could actually justify punishment and the discounting of people's interests. But as Strawson (1968) suggested, people often seem to think that the price of error theory about moral responsibility is error theory about *morality* – error theory, that is, about whether it is ever actually morally wrong to do anything. Thus Strawson characterized "pessimists" about free will and its relation to moral responsibility to be those who think that if determinism is true, "then the concepts of *moral obligation* and moral responsibility really have no application."

But as we had occasion to remark, the concept of moral wrongness requires responsibility in no sense stronger than any other notion that concerns our prospective reasons for action. To think we that we should do, think, or feel something, we must suppose that the thing is within the scope of causal influence of our normative thought, but no more. The fact that something is wrong entails that we should feel obligated not to perform it (unless we will omit it anyway); it does not entail that others are automatically justified in feeling anger at us if we don't do it. The very way we started worrying about whether there was such a thing as moral responsibility was our noticing that there is such a thing as *exculpation* – as an act's being wrong but not blameworthy. If, on account of the absence of Libertarian free will, we are never moral responsible, then all of our wrongdoing is simply thrown into the category of Gibbard's person who lashes out in a paroxysm of grief. It in no way threatens the fact what we do is wrong when we do it. The forward looking role of the notion of moral wrongness is to guide our attempts to avoid doing what we take to be wrong, just like the forward looking role of our judgments that we ought avoid our own pain is to guide our attempts to avoid pain. In neither case does it matter to the role of the concept in our reasoning if someone is justified in being angry at us and hurting us if we mess up in acting for those reasons.

I thus think that Strawson did us a bad disservice by lumping "feeling bound or obligated (the "sense of obligation")" and "feeling compunction" in with other things he called "reactive attitudes" like guilt and anger. Strawson initially started calling things 'reactive attitudes' because they were "reactions of offended parties and beneficiaries" and on account of their being "reactive attitudes *towards* good will or its absence." But feelings of obligation are not reactions to anything but our perceptions of our moral reasons, just as desires and other aversions are not reactions to anything other than our judgments about other kinds of reasons to do things out of them. Since we tend not to feel obligated to do things we are going to do anyway, we could, I suppose, say that these feelings are reactions to insufficient preexisting good will on our own parts, but that would be highly misleading in Strawson's context. Because it really doesn't matter to the role of feelings of obligation in responding to our judgments of their fittingness, or our judgments of what would be wrong to do, whether or not we are in any very strong sense responsible for anything. It would be as absurd to say "is it really fair to feel obligated not to do *X* given that this or that part of my psychology is beyond my control?" as it would be to say "is it really fair to be averse to pain given that this or that part of my psychology isn't up to me?"

The sense in which outrage or resentment is a "reaction" to preexisting wrongdoing is very different, for it is a reaction which is such that we can ask of it, "is that reaction really fair, given that this feature of that person's psychology was not up to her?" We already have a practice of admitting the wrong or the obligation – which is to say admitting the appropriateness of the feeling of obligation – and then going on to question the appropriateness of the angry reaction on account of diminished control. It is this practice, which itself presupposes the fittingness of the attitude of feeling obligated, that opens the way to the view that the more robustly "reactive" attitude of outrage is never appropriate. The most compelling argument for error theory about moral blameworthiness or responsibility thus relies upon, rather than in any way challenges, the fact that certain actions are morally wrongful and the fact that we have reason not to perform them on that account.

Chapter 5

Morality, Attitude Kinds, and the Honor System

In the last chapter we explored various of our moral concepts and how they are related to reasons for action. Along the way we have seen some important connections between our moral concepts. For one thing, I offered an analysis of supererogatory action that drew upon the notion of morally wrongful action. We have also been relying on the idea that all morally blameworthy action is morally wrongful, but as we emphasized, not all morally wrongful action is morally blameworthy (perhaps none is). But we might well wonder what, if anything, our moral concepts have to do with one another *qua* moral concepts. We saw an expansive sense of a "moral reason" as a consideration that contributes to the fittingness of a moral emotion, but what is it for an emotion to be "moral"?

It might not seem that the question of how to cleave off morality from other parts of ethics is very important. After all, we have explored various of the important concepts in terms of which we deliberate and guide our responses, and seen how they relate to what we have reason to do. We might happen to be inclined to call these concepts 'moral', but who really cares? One reason I think we should care is that understanding what our moral concepts have in common helps us to better understand certain relations of entailment between them. This enables us to see some important ways in which moral thought is distinct from but closely related to another kind of thought, namely thought about what is shameful, lowly, contemptible, and non-morally virtuous. Revealing how moral thought is related to this other kind of thought helps us to appreciate the practical import of the latter. It also helps us to see how the two kinds of thought could come together in certain ways, and understanding this can help defend our foregoing analyses

of moral concepts against a certain kind of worry about the universality of moral emotions.

5.1. Conceptual Connections Between Warrants for Distinct Moral Emotions

In chapters 2 and 4 we saw that there is a gap between something's being wrong and its moreover being blameworthy. When someone is exculpated, as in Gibbard's case of lashing out in grief, that person's conduct remains wrong but fails to be blameworthy. On the other hand, we have relied upon the idea that there is no such thing as conduct that is morally blameworthy but not morally wrong. Certain instances of conduct might, of course, be morally bad but not morally wrong – someone might think this, for instance, about selfishly failing to pay for an impoverished employee's medical expenses. But if it is merely morally bad then it befits only moral disesteem, or looking down on it with something like an impartial version of feeling screwed over. For conduct to be not only bad but also blameworthy it must also befit guilt on the part of the person who performs it and outrage or resentment on the part of others. But it seems that only wrongful actions could be candidates for such attitudes. An act can only befit anger and guilt after the fact if it was something that its author should have felt obligated to omit before the fact. But why exactly is this so?

This might actually be a place where a judgmentalist could seem to have an explanatory advantage. For a judgmentalist could offer the following explanation of why warrant for guilt and anger after the fact entail warrant for feeling obligated before the fact but not vice versa:

> "Feeling obligated to do *X* involves judging that it would be wrong to do *X*. On the other hand, feeling guilt for doing *X* and feeling angry at someone for having done *X* both involve judging that doing *X* was wrong *and* that you or the other person *responsibly* did wrong. Since an attitude's warrant is a matter of the correctness (or rationality) of the judgment it involves, the warrant for guilt and anger entail warrant for feeling obligated but not vice versa."

This judgmentalist account might seem to offer a nice explanation of another feature we have been presupposing too, namely the connection between warrant for guilt and warrant for anger. It looks like it is only fitting for others to be angry at me for having done things that I should also feel guilty for having done. How could you have a right to be angry at me for doing something I shouldn't even have to feel bad for doing myself? In the same way, it looks like I do not have to feel guilt for doing something unless it is the sort of thing that others are justified in being angry at me for doing. For if you have no right to be angry at me for doing something, why should I have to feel guilty for doing it? The foregoing judgmentalist account would explain the apparently necessary co-variation between the fittingness of guilt and the fittingness of anger: guilt is fitting on the part of the actor if and only if anger is fitting on the part of others because both of these emotions involve judgments of responsible wrongdoing and both of these emotions are fitting just when it is true (or rational to judge) that the actor has done responsible wrong.

Of course, as we have seen, it might be estimable for me to feel guilty for having done things that do not befit the anger of others, and it might even be disestimable for me not to feel such guilt, but this is distinct from my guilt's being fitting or warranted by my circumstances. Similarly, it might well be that I have done things for which I should feel guilt that you don't even know about. If you have no evidence of my crime (but I do), there may well be a sense in which it is fitting for me to feel guilt but it would be unfitting for you to feel anger at me for what I have done. But recall our distinction from chapter 2 between fittingness in the objective and subjective senses. If I have done something for which I should feel guilt in light of all the facts then it is objectively fitting for you to be angry at me, or fitting for you to be angry at me given the facts of the case, even if you do not know these facts. We should thus understand the connection between the fittingness of guilt and anger as first and foremost a connection between their fittingness in the objective sense. Secondarily we can understand this as a connection between the subjective fittingness of guilt and anger given the same set of evidence. That

is, if it is fitting for me to feel guilt for doing something in light of my evidence,[138] and you have the exact same evidence as I, then it is subjectively fitting, or fitting in light of your evidence, for you to feel angry at me for doing it.

Along these lines, we might also clarify the sense in which an act must be wrong if it is moreover blameworthy. In chapters 2-4, I gave analyses of wrongness in terms of the fittingness of feelings of obligation that were not explicit about whether the fittingness was objective or subjective. From the standpoint of the role that thoughts about wrongness play in our deliberations about what to do, the distinction does not matter so much. For while we strive not to do things that would be objectively wrongful, we can only be guided by our evidence, or our assessments of what would be wrong in light of that evidence or wrong in the subjective sense. But there really are two distinct senses in which feelings of obligation can be fitting, or mandatory, which correspond to two distinct senses in which we can appraise an act as morally wrongful. An act is subjectively wrongful, we might clarify, if it is fitting or mandatory to feel obligated not to do it in light of one's evidence. Thus if you have excellent evidence that someone is drowning, we might think that you should in the subjective sense feel obligated to jump in and save him, even if your evidence is misleading and he is actually just playing around or shooting a movie with the camera crew out of sight. Correspondingly, if it was only a matter of muddying your clothes, we might thus think that failing to try to save the swimmer would have been morally wrong in the subjective sense.

But, we would think, it would not be fitting in light of the facts of the case for you to feel obligated to try to save the swimmer, and correspondingly it would not have been wrong in the objective sense for you to fail to do so. Now you might wonder why we should care about wrongness in the objective sense at all. The reason, I think, is that it makes thinking about what to do a lot cleaner if we can separate it into two different components. The first component is a basic normative question about what to do if the facts of our case turn out to be this way or that. This we answer by means of the

[138] Even those who have perpetrated something for which they are culpable might lack the relevant evidence of their culpability. For instance, I might perpetrate a horrible crime and then suffer temporary amnesia. Less exotically, most of us are familiar with having woken up with pangs of guilt, unsure for a minute if we actually did something horrible in waking life or if it was only a dream.

reflective equilibrium methods that seek out the normative theory that makes best sense of our non-debunked intuitions about *what to do* in cases where we assume we know all the facts. So, a person might defend direct Utilitarianism as an answer to the basic normative question of what to do, where this theory is couched as one according to which we should do whatever of the acts available to us will bring about the most universal happiness. But the second component of inquiry into what we should do involves determining what our evidence actually supports, and how we should act in light of our evidence provided an answer to the basic normative question of what to do. Thus, whether or not we are direct utilitarians, we can ask what degrees of belief we should have about how various things will contribute to universal happiness, and we can ask what to do given our credences and given a view about what to do in the objective sense.[139] Answering these questions also requires basic methods of normative inquiry, but they are aimed at the question of *what to believe* given a certain set of evidence and what to do *given* (or assuming that one should have and pursue) certain credences and aims. By developing theories of objective wrongness, we can get on with the task of figuring out what we should aim at, and can leave to one side questions of what to believe and how to pursue our aims in light of our beliefs.[140]

But even if the notion of objective wrongness is in this way an important tool for normative theory construction, it is not the kind of wrongness to which our notion of moral blameworthiness is particularly related. This is clear enough when we consider my

---

[139] One question along these last lines, which is distinct from both what we should aim to achieve and what degrees of belief we should have about various things is whether we should, like evidential decision theory says, perform the act that has the highest expectation of being an act that achieves what we have most reason to achieve, or whether we should we rather, like causal decision theory says, perform the act that has the highest expectation of bringing about what we have most reason to achieve.

[140] Thus, criticisms of direct consequentialism on the grounds of its being "too hard to figure out what will bring about the best consequences" are not playing by the rules of this partition of normative questions, and those who make these criticisms should be shown the errors of their ways.

There are, I suppose, ways of making something like this kind of criticism that are still playing by the rules. The idea behind these would be that we have some independent evidence that we can know what we have moral reason to do in our actual circumstances without having to calculate even its *rationally expected* consequences. But in practice this probably does little more than reiterate whatever independent intuitive evidence there is against direct consequentialism – e.g. that we don't need to know the expected consequences of promise keeping to know that we have *some* moral reason to keep promises. (Incidentally, the view that we can know what we have *most* moral reason to do in our circumstances without calculating the expected consequences of our actions is wildly implausible. Can we really know that we have moral reason to do *X* without having some view as to whether or not *X* will cause all sentient beings to be tortured for all eternity?)

doing something that was wrong in the objective sense but not in the subjective sense. For suppose that I am walking down the street to the library, and all of my evidence tells me that taking the next step will have no effect save propelling me closer to where I am trying to go. But suppose that my evidence is misleading: taking the next step actually sets off a trigger buried beneath the ground that causes a school full of children to explode and kill them all. We will presumably have some kind of objective moral theory according to which my taking the step was wrong in the objective sense (for surely we think that it is wrong to painfully kill people with lives worth living for no reason). Before the fact it was objectively fitting for me to feel obligated not to do what I did, in the sense that had I been appraised of all the facts of my situation this is how I should (in the subjective sense) have felt. But since I had no way of knowing that taking the next step would harm the children, it was not subjectively wrong - or fitting for me to feel obligated not to do what I did given my evidence. As such, it surely would not be fitting for others to feel angry at me for taking my step (indeed in the *objective* sense of how they should feel in light of all the facts). For again, how could I have known that I was harming the children? In the same way, while I might well feel guilt for having stepped on and set off the bomb, this guilt would not be fitting or justified (though, given my causal relation to what happened, I suppose that some people might think it estimable for me to feel guilt, or even disestimable or worrisome for me not to).

Thus, for an act to be blameworthy or befitting of guilt and anger, it must be that the action was something that it was wrong in the subjective sense to do. It also seems that, to the extent that someone is sufficiently morally responsible for her actions, subjective wrongness may be sufficient for blameworthiness even in the absence of objective wrongness. This is presumably how we think about the way in which attempted murder is blameworthy and deserving of punishment. If I had excellent evidence that my gun was loaded with real bullets when I put it to Jones's head and pulled the trigger, then even if unbeknownst to me someone replaced the real bullets with blanks, we will tend to think that (if I am morally responsible for my action) I have done something blameworthy. We will tend to think, that is, that I should feel guilt for what I have done and that others are justified in feeling angry at me for doing it.

Does this show that objective wrongness is completely irrelevant to culpability? One barrier to thinking this might seem to be that we tend to think that, even holding fixed the evidence and mental states of the perpetrator, murder is more culpable than attempted murder. The standard thought seems to be that you should feel guilty for trying to kill people and we are justified in being angry for your trying to do so, but that you should feel even more guilty if you actually succeed in killing someone and we are justified in being even more angry at you if you do. One way to explain this might be to say that while subjective wrongness is a necessary condition for culpability, one subjectively wrong act can be more culpable than another depending upon how objectively wrong it was. But another way to go might be to maintain that subjective wrongness is all that matters to culpability but allow factors beyond our evidential situation to determine which acts it is that are assessable as subjectively wrong. Thus, since in cases of attempted murder there is no actual murder, there is no actual murder that we can appraise as having been subjectively wrong – the only thing we have to appraise as subjectively wrong was the attempt. On the other hand, in cases of successful murder, we have both an attempt and a successful murder to appraise as subjectively wrong, and we can say that murder is more deeply subjectively wrong than attempted murder. I am not sure if either of these two approaches is superior to the other, but for our purposes it will not matter if there is. The important thing to bear in mind is that culpability requires wrongness in the subjective sense.

Apart from the dependence of the warrant of guilt and anger on each other and on feelings of obligation before the fact, are there any other connections between warrants for moral emotions? We might think that an action can only be morally estimable if it is also morally permissible in the subjective sense. We might wonder, though, if it is coherent to think that there is such a thing as admirable immorality in the sense of action that is subjectively wrong but still morally estimable. Suppose, for instance, that you can make a huge contribution to a war effort by staying comfortably behind the lines and working as a translator, but that some of your closest comrades have been captured and will be executed unless you risk almost certain death on a daring mission to rescue them. It might be no easy thing to weigh your moral reasons to stay back to contribute to the

wider war effort against your moral reasons to save your friends. But suppose we judge that all things considered it is morally obligatory to stay behind: given how much you can do for the war effort, it would be wrong for you to risk depriving headquarters of your services by getting yourself killed. It might seem coherent for us to think this but to go on to judge that, should you go on to risk your life for your friends, you your action will be morally estimable. Of course, we might think that if it really is wrong given your evidence to risk depriving headquarters of a translator, it would be a sign of weakness or indulgence rather than moral virtue to go on the rescue mission. But the point is that it might not be incoherent to think otherwise.

A less contentious connection might be one between moral goodness and supererogation. As we saw, it seems coherent to think that some acts are morally good but not supererogatory (e.g. putting up with the maltreatments of someone you are carrying to safety), and it seems coherent to think that some acts are supererogatory but not morally good (e.g. someone's saving your life at the cost of his out of an irrational belief that he needs to do so to avoid hell). But this much might seem to be right: if a person does something that, given his evidence, is supererogatory, and he does it for the reasons that make it supererogatory, then he does something that is morally good. Thus, suppose that your evidence is such that you are in Bugsy's situation – where you will be shot in the head if you save the child from drowning – and you save the child because you take her need for help and your ability to provide it to count in favor of feeling obligated to do so. It seems that if we agree that your act was supererogatory for the reasons on which you acted, or that (1) you didn't have to feel obligated either to save the child or to omit saving her, (2) you really did have reason to feel obligated to save her for the reasons that you thought, and (3) you had strictly more reason to feel obligated to save the child than to do something else that you could have permissibly done (like sit tight), then we should agree that you did something morally estimable, or something that merits moral esteem.

In the same way, we might think that there is a kind of connection between morally disestimable action and action that one should feel obligated not to perform. As we have seen, it seems coherent to think that some actions, like failing to help an impoverished employee with medical expenses, might be morally disestimable but not

morally wrongful (in both the objective and subjective senses). If so, then an act can be disestimable without one's having to feel obligated not to perform it. But it might seem that if it really is disestimable, there must be at least some reason to feel obligated not to perform it, even if this reason is insufficient to make the feelings of obligation mandatory.

In light of the full range of connections there might seem to be between reasons to feel our various moral emotions, the judgmentalist account of the connections between the fittingness of guilt, anger, and feelings of obligation might not look so attractive. Because the judgmentalist seeks to explain the fittingness of an emotion in terms of the correctness or rationality of the judgment it involves, she seems to have no clear way of making sense of how some emotions might be rationally optional rather than rationally mandatory. If reasons for emotion are reasons to hold the judgments that they involve, it would seem that all emotions would be just as mandatory as the moral beliefs they involve. As such it is unclear how the judgmentalist might relate the fittingness of feelings of moral esteem and disesteem to reasons to feel obligated to do things. She would, one supposes, want to say that moral esteem and disesteem involve judgments of moral goodness and moral badness, and that these are somehow conceptually related to judgments of moral wrongness, but it is unclear how she will explain these connections without speaking antecedently of conceptual connections between reasons to have attitudes like moral esteem, moral disesteem, and feelings of obligation (as only we anti-judgmentalists can do informatively).

There is at least one other important problem with the judgmentalist's attempt to explain the conceptual connections between the fittingness of guilt, anger, and feelings of obligation in terms of the judgments that they allegedly involve. The first is closely related to the general criticism of the judgmentalist account of attitude regulation in chapter 2. This is that, when we think that guilt and anger are fitting only if wrong has been done, we do not simply mean that it is unwarranted to make judgments of responsible wrongdoing if it is unwarranted to make judgments of wrongdoing. We mean, moreover, that it is unfitting to have the affect, motivations, and tendencies to direct attention of guilt and anger after the fact if it is unfitting to have the affect,

motivations, and tendencies to direct attention of feelings of obligation before the fact. It is not as though, if I have done nothing wrong, you are justified in having deep negative affect at me, feeling like yelling at me and beating me up, plotting your revenge on me, and so on, as long as you please just don't call what you're feeling 'anger', 'outrage', or 'resentment'. Similarly, if I have done nothing wrong, it is not as though it is fine that I feel horrible about myself, dwell obsessively on what I did, try to restore things to the way they were before, or feel like submitting myself to the death penalty, so as long as I just please don't call it 'guilt'.

There is thus a dependence of warrants for the conations - the affect, motivation, and directed attention – associated with guilt and anger upon each other and upon the warrant of the conations associated with feelings of obligation. But why on earth would any of this follow from mere relations of warrant between the judgments needed to "brand" these conative syndromes as ones of guilt and feelings of obligation respectively? The judgmentalist seeks only to explain:

(JD) The judgment that "brands" a syndrome of affect, motivation, and attention direction as one of guilt or anger is warranted only if the judgment that "brands" a syndrome of affect, motivation, and attention direction as one of feeling obligated is warranted

But what we really want explained is:

(CD) The syndrome of affect, motivation, and attention direction associated with guilt is *itself* warranted just in case the syndrome of affect, motivation, and attention direction associated with anger is, and both of these are warranted after the fact only if the syndrome of affect, motivation, and attention direction associated with feeling obligated is itself warranted before the fact.

But (JD) seems to be of absolutely no use in explaining (CD).

An additional advantage of eschewing the judgmentalist's approach to explaining the dependence of the fittingness of guilt and anger on that of feeling obligated concerns how we are to make sense of the concept of moral responsibility. The judgmentalist's

explanation appealed explicitly to the notion of moral responsibility – on her account, the fittingness of guilt and anger before the fact entail that of feeling obligated after the fact because the former involve judgments of responsible wrongness and the latter involve judgments of wrongness. But what is it to judge that someone is morally responsible for what she does? We might think that it is to make a special kind of normative judgment that itself calls out for explanation. At the end of the last chapter I surveyed some ways in which we might extend our everyday thinking about responsibility to get to the conclusion that no one is morally responsible. But what exactly was I doing? Perhaps I was surveying an explicitly normative argument – the kind of good old fashioned attempt to best unify non-debunked normative intuitions that we keep talking about. Perhaps the argument was in support of the substantive view that we are not morally responsible unless we have a certain kind of control of our mental states that we do not seem to have. In the same way, those who hold opposing views, like the Barbarists I mentioned, might have equally coherent (though perhaps not equally plausible) views. But if this is so, what were the Control Theory that the argument sought to support, and the Barbarist theory it sought to refute, really theories about?

One attractive answer might just be that they were theories about the gap between what befits feelings of obligation before the fact and what befits feelings of guilt and anger after the fact. Thus the Control Theory amounts to the view that it is not fitting to feel angry at someone who failed to do what she should have felt obligated to do if some of the mental states that led her to do it were things about her that were beyond her control. Versions of the Barbarist theory, on the other hand, would hold that it is fitting to be angry at a person who failed to do what she should have felt obligated to do even if she did it as a result of mental states outside of her control. (This view was called Barbarism, recall, because given the fact that anger involves intrinsic motivation to punish and the Warrant Composition and Rational Ends Principles, it entails that it is worth seeking to punish people who do things as a result of factors beyond their control as an end in itself.)

In conjunction with our picture of judgments of fittingness as basic kinds of warrant assessments that govern valenced attitudes *qua* syndromes of affect, motivation, and attentional focus, this understanding of what's at issue between the control theorist

and the Barbarist might seem to explain quite a lot. For in situations in which the Control Theorist takes it that a person failed to do what she should have felt obligated to do but lacked the appropriate kind of control (which, *pace* Libertarianism, will include all actual situations), she will have a propensity not to anger at her and manifest symptoms of recalcitrance if she does anger. In particular, she will (if she recognizes an entailment of her view) not take it that there are considerations that justify harming the person as an end in itself, and she will struggle against any recalcitrant motivations to do so. The Barbarist, on the other hand, will have no such qualms about anger and aggression towards those who fail to do what they should have felt obligated to do even if she recognizes that they lacked the kinds of control in question. She will not have a propensity to abstain from such anger, she will not manifest symptoms of recalcitrance if she has it, and she will (if she recognizes an entailment of her view) take it that there are reasons in favor of punishing the people in question as an end in itself.

But if we are to eschew the judgmentalist explanation of the relationship between warrants for our various moral emotions, how are we to explain these relationships? As we have tacitly seen, the connections between the warrants are not merely substantive but conceptual. There seems to be something incoherent about thinking that something is blameworthy but not (subjectively) wrong, something incoherent about thinking that our anger at someone is fitting but that she should not feel the least bit guilty. Now there seems to be nothing about the general notion of FITTINGNESS or WARRANT which would explain why there are conceptual connections between warrants for distinct moral emotions. So the only thing left would seem to be that there is something about our moral emotion concepts, our concepts of GUILT, OUTRAGE, FEELING OBLIGATED, and so on, which is not something about their involving related moral judgments, which explains how their warrants are related.

An idea at which I hinted earlier is that even if we are not judgmentalists, our emotion concepts might encode more than just the idea of a certain affective, motivational, and attention directing syndrome. Of course, I have been arguing that this is the feature of our emotions concepts that is primary, and this part of what our emotion

concepts pick out is what is governed by our assessments of warrant. But there would be no barrier in principle to our emotion concepts making reference to each other through the idea of one being warranted only when the other is. Thus, our notion of GUILT might not just be the idea of a state that involves such and so negative affect, motivations to make amends, and tendencies to dwell on what one has done, but such a state, which, by the way, is warranted only if some other state is.

There is, I think, independent intuitive support for the idea that our moral emotion concepts work this way. Suppose that there was a race of being from another planet – Twin Earth of course – who had thoughts about states with the same affect, motivation, and attentional focus as those of guilt, outrage, and feelings of obligation, and who thought that each was sometimes warranted. Let us call the concepts in terms of which the Twin Earthians think about these states TWILT, TWOUTRAGE, and FEELING TWOBLIGATED. The Twin Earthians have thoughts to the effect that certain things, like physically assaulting someone, warrant twilt after the fact, and as a result have a propensity to feel bad about themselves when they assault someone, dwell on what they did, and do nice things for the person they assaulted. They also have thoughts to the effect that they are justified in being twoutraged at each other for doing certain things, like failing to save each others lives. When Twin Earthian fails to help another, the others tend to have a negative affective state towards her, dwell on the fact of her not saving, are intrinsically motivated to harm her, and so on. And Twin Earthians think that they should feel twobligated to do certain things, like keep their paws off each others' property, and as a result go into a negative affective state when they are tempted to take each others' stuff, which involves being intrinsically motivated not to do so take it.

What is noteworthy about the Twin Earthians, however, is that they do not think that the things that warrant any one of twilt, twoutrage, or feelings of twobligation warrant the other two. For instance, they will deny that it makes any sense to feel twobligated not to assault each other. If you ask them why they don't assault each other, they will probably tell you something like, "Well, if you don't mind having to feel twilty afterwards and spend all of this time having to be nice to someone it's not so bad. It *is* interesting to watch a person suffer, but apart from not wanting to be in a position where you have to feel twilty (which kind of sucks and is kind of annoying), you probably don't

want to get a reputation for beating people up, because then everyone will run away from you and you won't get anything done."

If you ask the Twin Earthians what happens if they fail to save someone from drowning, you will be told something like, "It totally sucks because then people really start coming after you. They aren't just interested in seeing you suffer for fun, they are actually twoutraged at you. What's more, the people who don't feel twoutraged know they probably should, and they'll make it a point to try to come after you too, even if they think it'd be more fun to beat up someone else instead. Everyone knows that they're going to have to feel twilty for beating you up, but they're so usually so twoutraged at you or so determined to get you that they usually think it's totally worth it. We have a saying: 'If you fail to save someone from drowning, you'll wish you beat someone up, because having to feel twilty is nothing compared to getting beaten up by that wolf-pack that comes after you!' It *is* kind of fun to watch someone drown, but a lot of the time people save someone who is drowning just because they're so afraid of getting beaten up later. If you see someone just stand by and watch someone else drown, you can be pretty sure that that guy's either really brave or really stupid."

Finally, if you ask them what happens if someone steals something, you might get something like, "Yeah, I don't understand why some people steal. I mean how *can* you steal, it's *their* stuff; you can't just take it! I remember this one time I stole the only food this one guy had and ate it. He was really disappointed. He asked 'Didn't you feel twobligated not to steal my food?' and I said 'Yeah, I don't know what was wrong with me; it totally wasn't worth it.' He asked me if I was fully responsible, and I said "yeah, if I had beat someone up I'd totally feel twilty, and if I'd let someone drown you'd totally be justified in being twoutraged at me. I guess I just messed up. I'm so embarrassed." He said 'Well, don't worry, everyone makes mistakes. This is just rotten luck for me, because I'm going to starve now.' And I said 'Haha, yeah, I know. Well, I'd better get going, I've got a whole bunch of food at home. Best of luck surviving.' And we've been pretty good friends ever since."

If life on Twin Earth is as described above, there is no doubt that it is quite radically different from life on earth. In particular, it does not seem to me that the Twin Earthians' concepts, which they express with 'twilt' 'twoutrage' and 'feeling

twobligated' are the same as our concepts of GUILT, OUTRAGE, and FEELING OBLIGATED. Even though they are thinking about attitudes with the same affect, motivations, and attentional focus as these, the Twin Earthians they do not seem to be thinking in terms of our emotion concepts. A natural suggestion for why this is so is that to have thoughts involving the concepts of GUILT, OUTRAGE, and FEELINGS OF OBLIGATION, you have to think that instances of the first two concepts are warranted only when an instance of the other one is, and that instances of both are warranted only when a corresponding instance of the third is.

The suggestion, then, is that views about when a conative syndrome is warranted can enter into the folk emotion concept by which we pick out the syndrome. We might also think in this regard about why the Contour Thesis of the last chapter true or what guarantees its truth. A suggestion along the present lines would be that it is part of our very notion of feeling obligated not to do something, or having the sort of guilt-tinged aversion that feels like one "just can't" do it, that this attitude is rationally mandatory only if it is rationally mandatory to be most strongly motivated to do what it motivates us to do. I believe that considerations similar to the above twin-earth thought experiment could be adduced in favor of this view. Thus, if we had people who insisted that in some circumstances it is rationally mandatory – and not just very disestimable not to - feel twobligated to do things that it is not rationally mandatory to be most strongly motivated to do, or rationally permissible not to do, I think that there is reason to think that they would not be tokening the concept FEELING OBLIGATED.

5.2. A Speculative Evolutionary Story

It might appear to be rather surprising, however, as to why we would have moral emotion concepts that work in the way just described. Why would views about what kinds of states are warranted when others are be showing up in our very ideas of such emotions as guilt, outrage, and feeling obligated? I think that considerations of how our moral emotion concepts likely emerged would be helpful in answering such a question. In particular, I think that evolutionary considerations suggest that something about

interrelated warrants would have run deep in our emotional lives. So deep, in fact, as to show up in the clauses of the folk psychological theory from which we extracted our emotion concepts. Given the coordinative role of moral emotions, there would have been reason to take their warrants to be interrelated whatever else we took them to be. And given these pressures, there would have been a great deal of predictive value in attributing these emotions to people on the assumption that they would take their warrants to be related in these ways.

Let us begin first with an account of how the conative syndromes that constitute our moral emotions would have come to be causally interrelated, quite independent of judgments about their fittingness. Stories of how moral emotions help to facilitate coordination have been offered by many, including Trivers (1971), Gibbard (1990), and Sanfrey et al (2003). I shall begin with speculation that will be most germane to the understanding of moral emotions we have been developing so far. Let us consider the metaphor of genes that inhabit early hominid organisms. The metaphorical objective of each gene is to pass down as many copies of itself to future generations as possible.[141]

Suppose, then, that you take up the metaphorical perspective of a gene located in a hominid organism that is about to interact with another gene located in another such organism. You and the other gene have the metaphorical objectives of optimizing your representation in future generations, and face you face the coordination problem known as the Prisoners' Dilemma [PD]. When your organism meets that of the other gene, each organism can behave in a way towards the other that is intuitively nice, which we shall call "cooperating," it can do something that is intuitively nasty, which we shall call "defecting." What matters to you as a gene is the fact that if the other organism does something nice to you (e.g. doesn't take your organism's food), that helps you make copies of yourself, while if your organism does something nasty to the other (e.g. take its

[141] I think that applications of evolutionary game theory to biology are in general best thought of as of "battle-bot" scenarios in this kind of way, where we employ the metaphor of genes seeking to maximize their representation in future populations and employing strategies to do so. Not only is this metaphor more accurate than that of organisms trying to employ such strategies, but it better discourages crude conflations of genes' metaphorical "rationality" and "interests" in maximizing their future copies with the preferences, welfare, or reasons of the beings that such processes cause to exist. This is particularly important for understanding the evolution of emotions – which involve intrinsic motivations on the part of organisms to do various things – and normative governance – which is a mechanism within an organism that guides attitudes in certain ways but does not in any way represent the objectives of genes.

food), that helps you make copies of yourself too. We may illustrate the expected payoffs to you and the other gene depending upon what your organisms do as follows:

		Rival Gene	
		Rival's Organism Cooperates	Rival's Organism Defects
You	Your Organism Cooperates	2, 2	-4, 4
	Your Organism Defects	4, -4	-1, -1

Figure 7: Prisoners' Dilemma for Genes

In this table the numbers represent payoffs to you (the first entry in each box) and your rival (the second entry) in terms of increases in expected of representation in future generations. If you know that your organism is only going to face this situation with the same rival once, then it pays to program it to defect, for whether the other organism defects or cooperates, you will do better defecting (4 vs. 2 and -1 vs. -4). But what if organisms in fact interact with each other in iterated sequences of these kinds of situations many times over the course of their lives, and each gene can tailor its organism's action on the next round of play in response to what it and the other organism has done on previous rounds?

We can expect this kind of environment to characterize that of many social animals, including perspicuously our primate ancestors.[142] This was essentially the question posed by Axelrod and Hamilton (1981), who ran a computer competition of iterated-PD strategies by pairing them (and a completely random strategy) against each other, playing for 200 moves, and determining the winner by summing the strategy's payoffs across different partners. The winner, both of the first round of competition and another round with entries submitted after looking at the results of the first round, was the

[142] In fact Kitcher (1993, 1998) has some suggestions as to how this kind of reasoning might be able to be used to explain the very emergence of such sociality among organisms where none (or at any rate much less) existed before.

simple and elegant TIT FOR TAT. TIT FOR TAT is a strategy that cooperates on the first round, then on each subsequent round cooperates if its opponent has cooperated on the previous round and defects if its opponent has defected on the previous round. Axelrod and Hamilton attributed TIT FOR TAT's strength to three of its central features: it is generous on the first round, provokable (it defects if its opponent defects), and forgiving (is ready to cooperate after one round of retaliation).

So suppose, then, that due to its adaptive benefits, you decide to program your early hominid organism to play TIT FOR TAT. Now you do not get to start from scratch – your organism already has powerful motivations to do the mean, defective thing that gets you a short run kick of reproductive advantage at the expense of the other organism. So you need to program in some mechanism that will cause your organism to restrain itself. Perhaps in principle you dream of giving your organism strange new desires and massive computational capacities so that it would become literally the way you are metaphorically: desirous of increasing your representation in future generations, and decided on playing TIT FOR TAT as the best way of doing this. But alas, you cannot; it will be millions of years before things like your organism even come to have the concept of a gene. So a more realistic option is to give your organism a new intrinsic motivation not to defect, which can do battle with its existing motivations to do so. A conative syndrome like that constituted by the prospective guilt-tinged aversion of feeling obligated not to defect might be ideal. Its negative affect and attention direction will alert the organism to the importance of the situation and aid the efficacy of its motivation not to defect.

But that's only the first ingredient you'll need to get your organism to successfully implement TIT FOR TAT. You also need a motivation that will cause it to punitively defect if the organism with which it's playing defects first. Again, you could start from scratch, but the practicable option here really seems again to program a motivational syndrome – one that motivates one's organism to behave aggressively towards another organism in the face of its defection. This looks an awful lot like another familiar motivational syndrome, namely resentment.

Now following the first competition of Axelrod and Hamilton (1981), there was further work on optimal strategies in iterated Prisoners' Dilemma games. Some, like that

of Nowack, May, & Sigmund (1995), suggests that variants on TIT FOR TAT (like Generous TIT FOR TAT and PAVLOV) can actually improve on its evolutionary stability in what we might consider more realistic environments with some degree of randomness. Krebs (2005) characterizes these strategies as "kinder, gentler, [and] more forgiving," and suggests that they can "reestablish the string of mutually beneficial cooperative exchanges, which can be accomplished either by the selfish player making up for his mistake or the victim giving the selfish player a second chance." As a gene trying to implement these kinds of strategies, you would want to implement a psychology in your organism that would motivate it to "make up for its mistake" in order to restore cooperative exchanges with other organisms playing similar strategies if randomly its defecting impulses got the better of it. A good bet for this would be to instill in it a "placating, self-punishing emotional response," felt when your organism has defected and its partner stands to defect punitively. This is precisely what Gibbard (1990) describes as the role-description for our emotion of guilt in placating anger. So we here see similarly that, given what would most effectively enable the genes to pass themselves down in the presence of these coordination games, they will program their organisms to feel guilt for having done something only if it was the kind of thing they'd try to deter it from doing in the first place with feelings of obligation not to defect without provocation. Further, with randomness introduced we might be able to see why it would be *only* but not *all* such defections that the genes would see fit to punish with resentment and correspondingly have to seek to placate with guilt. It makes sense to try to "forgive" organisms for relevantly "random" defections that over-power the deterrence of feelings of obligation not to defect.

To get much of a coordinative society among our primate ancestors going in the first place, the genes in our ancestral populations would probably have had to go through something like the above evolutionary processes, instilling our ancestors with feelings of prospective guilt-tinged aversion, resentment, and guilt. Equipped with these, they could be able to form close enough bonds with non-kin in order to reap important new benefits from multi-lateral coordination made possible by mutual restraint (cf. e.g. Kitcher 1998). Unfortunately, this also opens up new scope for genes to gain reproductive advantages at the expense of other genes (and organisms to gain resource advantages at the expense of

other organisms) by programming their organisms to take the benefits of this multi-lateral cooperation without doing their bits to contribute to society (even if they still restrain themselves in bi-lateral exchanges).

Now, in the case of personal interactions among organisms, the foregoing could explain how feelings of obligation not to defect against a particular other and punishment via resentment could keep defectors in line. But as Krebs (2005) observes:

> Although detecting and punishing those who cheat you personally may pay off better than ignoring them, the costs of taking it upon yourself to catch and punish free-riders who fail to contribute their share to society usually outweigh the gains. Better to let someone else do the dirty work.

Boyd and Richerson (1992) have suggested that this problem can be overcome by punishing members of groups who failed to punish free riders, and Gintis, Bowles, Boyd, Fehr (2003) showed how a mechanism that disposed individuals to cooperate and to punish those who failed to cooperate could invade a population of selfish individuals and become evolutionarily stable. Perhaps most importantly, Krebs (2005) notes:

> Price, Cosmides, and Tooby (2002) adduced experimental evidence that two motivational systems have evolved to overcome the free-rider problem. One disposes people to punish free riders, and the other disposes people to recruit cooperators by rewarding cooperation.

These motivational system mentioned here seems to play the exact roles of our moral emotions of outrage (or anger directed towards those who defect against others, though not on behalf of a particularly offended party either) and moral esteem. In the presence of outrage the same rationale that explained why guilt would tend to be felt so as to placate resentment will explain why guilt would tend to be felt in order to placate outrage. With these mechanisms in place we can also see that it would make more sense for the genes now to program their organisms to pay their fair share – and avoid punishment and costly amends – by feeling prospective guilt-tinged aversion towards engaging in social defection in the first place. But again, we must recall that against a more realistic environment with randomness it will make sense for the genes to punish *only* but not *all* defections with outrage - it makes sense to try to "forgive" organisms that are subject to relevantly "random" episodes of impulses to defect over-powering the

deterring motive of feeling obligated not to. Hence, we can see that the logic of the decision problem faced by the genes of our ancestral hominid populations might cause them to coordinate to program us to feel guilt, outrage, and resentment only towards those acts towards which they would try to deter with prospective guilt-tinged aversion in the first place, and why they would program an organism to feel guilt when and only when they programmed others to feel outrage or resentment.

It is worth noting that if this is how feelings of guilt, anger, and obligation came to be related, the most central feature of them would have been that they were coordinated in these ways. Whatever varied and different things filled in the blanks of counting as "cooperating" and "defecting," it would make sense to deter the organism from defecting with feelings of obligation, to punitively defect only towards the unprovoked defections of others, and to placate the punitive defections with restorative cooperation.

Now at some point in evolutionary history, we came to have not simply syndromes of conation like moral emotions but judgments of fittingness or warrant that governed those syndromes. Gibbard (1990) argues convincingly that the adaptive function of these judgments of warrant had much to do with enabling our moral emotions to play their cooperation-enhancing roles, but more flexibly and in a way better suited to complex and constantly changing social environments. Hard-wiring what shall count for emotional purposes as "cooperation" and "defection" might work well enough in relatively stable environments. Perhaps something like this is what has been implemented among our primate cousins, who manifest what appear to be relatively fixed patterns of proto-versions of our moral emotions (see for instance de Waal 1991, 1996). However, as Quartz and Sejnowski (2002) observe, the ancestral hominid environment that led to the adaptation of most complex neo-cortical developments (including, it would seem, judgments of the fittingness of moral emotions) took place in an environment of almost unprecedented variability and rapid climatic change. During these populations the size of hominid groups increased dramatically, and this together with the other accompaniments of neo-cortical development led to a great deal of social variation and complexity in addition to that more directly occasioned by the novel non-social

environmental features. Against a background of such complexity and variation, hard-wiring in what would count as "cooperating" and "defecting" would be hopeless. Organisms would have to be equipped with mechanisms for determining that for themselves, and the mechanism selected was that of our judgments of the fittingness of moral emotions with their direct propensity to guide them.

But in order for our judgments of warrant for moral emotions to play the adaptive role of facilitating coordination more flexibly, each hominid would have had to assume that an act warranted either guilt or anger after the fact only if it warranted the other, and only if its author should have felt obligated not to perform the act before the fact. For the evolutionary function of our having views about when moral emotions were warranted was to, as it were, coordinate on which things we were going to enforce cooperation by means of our moral emotional mechanisms. There would have been no adaptive benefit in judging guilt or anger fitting towards things other than those we would think it fitting to avoid out of feelings of obligation in the first place. Otherwise the punishment and placation would be adaptively pointless - from the standpoint of cooperation it only pays to enforce sanctions against things that it pays to deter in the first place. There also would have been little adaptive benefit it judging either guilt or anger fitting towards different things. Anger could play a deterrent role in the absence of guilt, but once guilt is on the scene it would certainly seem to minimize social friction and wasteful conflict for it to be felt towards the same things as anger so as to placate it – this would be the next best thing to preventing these things in the first place with feelings of obligation.

It may also be worth noting that there are reasons to think that, in order for feelings of obligation to play their primary role in deterring prospective defection, we would have had to think that we had most reason to be swayed in their direction whenever we thought them mandatory. Kitcher (1998, esp 299-303) discusses how fragile cooperation can be when inhibitions against defection are just one motive in the "internal melee" among many, as with our evolutionary cousins, the chimpanzees. He suggests that normative control helped evolve as a way of more reliably preventing prospective defection, thus ensuring social stability, liberating social effort from the need to constantly break and mend social bonds, and enabling cooperation in larger social units. But none of this would be possible if the fact that an action would constitute a

defection was just one consideration that we weighted in our reasoning among many. For that would just replicate the problem of fragile cooperation; just as defection is not reliably deterred among chimpanzees because they are "wantons" (they do not, that is, normatively govern their emotions and intentions), so too it would not be reliably deterred among us normatively governed beings if our normative governance did not come down emphatically on the side of non-defection.

Of course, there would have been situations in which "non-cooperation" would have paid big, and as a gene programming a hominid you would have wanted to take advantage of this. But since your hominid was apt to be a simple soul (and even the most complex souls are too easily prone to self-serving rationalization), it might have been best to pursue the gains from non-cooperation in the following two ways. First, the fact of imperfect normative governance would here provide you with a kind of kludge – even if your hominid's normative governance comes down against defection, it can still defect out of weakness of will, and we can expect *akratic* defection to be more likely the more your organism stands to gain from it. Second, if the gains from non-cooperation are important enough, you might have your hominid stop thinking that this kind of non-cooperation is aptly treated as defection. After all, the other genes are all in the same boat with their hominids, and all you genes might be far better off if you don't get carried away in what you try to make each others' hominids do in the service of social cooperation. The iterated PD-setup assumes that the gains to each gene from both hominids cooperating are greater than the gains to each gene from both hominids defecting. But if, for instance, we counted as "cooperation" things like your hominid trying to save someone else by engaging in unarmed combat with a bear or saber-tooth cat, cooperation might well pay less than defection. It would not hurt to let the cooperative emotions weigh in favor of doing things like that; after all, if there are mechanisms that reward cooperators this might be a good idea. But what would seem important would be enabling your hominid to distinguish emotionally between situations like this and situations around which it makes evolutionary sense to demand cooperation.

Remember, however, that much of the point of this normative governance business is flexibility. You don't get to program in emotional proclivities that say that particular things are definitely in or that particular things are definitely out, because the

things in question and their adaptive significance are likely to change in a few decades or hundreds of years, and there is no way that you as a gene can keep pace. What you can do, however, is make sure that your hominid rules out performing whatever sorts of things it does treat as defections, in the sense of acts which are apt for what Mill called "the sanctions of society and conscience" (unless they are excused). To do this you can have its system of normative control mandate, rather than simply allow, feeling obligated not to do these things (unless, of course, it is going to do them already), and to mandate that motivation in this direction be strongest. The motivations might not listen, but there is only so much you can do by way of tinkering with your organism's psychology over a short span of time.[143] The requirement for feeling obligated to omit something can thus mark it before the fact of its performance as the sort of thing that it is crucially important to avoid doing, the sort of thing that it is worth punishing after the fact in the absence of excuse. Feelings of obligation are a helpful force of social motivation that it would be unwise to rule out in other circumstances, but if we are to avoid the problem of unreliable deterrence we had better have a way to distinguish between the cases in which feelings of obligation are good to throw into the mix and cases in which motivation had better rally to their side. A distinction between thoughts about rational mandates for feeling obligated and rational options for feeling obligated could be just the ticket.

5.3. Arriving at our Folk Emotion Concepts

If the above speculation is on the right track, then we could expect the psychology of our hominid ancestors to have been characterized by two features. First, they would have come to have coordinated tendencies to have the conative syndromes of affect, motivation, and attentional focus associated with guilt, resentment / outrage, and feeling obligated. They would have tended to feel the first two towards actions after the fact of their performance only in the presence of each other and only when the act's author would have felt the last after its performance. Second, these coordinated tendencies

[143] And anyway, as we saw, this sort of *akratic* defection is a useful kludge for reaping occasional rewards of defection.

would have been flexibly extended to all kinds of novel social and environmental situations through normative views according to which these conative syndromes were warranted towards the same things. That is, our ancestors would have thought an act to warrant guilt and resentment / outrage after the fact only if its author should have felt obligated not to perform it before the fact. I now want to turn to what bearing this might have on the ordinary emotion concepts we came to have.

In addition to states like beliefs and desires, emotions and judgments, our ancestors at some point came to have beliefs about these states. They came, more generally, to have a theory about the kinds of inner states people are in, which lead people to behave as they do in the circumstances in which they find themselves. This is the kind of theory we still use today when we engage in our everyday thought and talk about beliefs, desires, judgments, and emotions. When we explain someone's behavior as the result of what she believed, desired, felt obligated to do, felt angry about, intended, and so on, it is in the terms of this theory that we explain her behavior. We are relying on a theory that says things like 'people tend to do what they intend to do', 'people tend to form intentions to do what will realize the objects of their desires and motivations according to their beliefs', 'people tend to feel guilty for doing what they think is wrong, and this motivates them to try to make amends for it', and so on. It might sound odd and unfamiliar to say that these things we usually presuppose or tacitly believe are a 'theory'. For one thing we are used to thinking of theories as things that someone comes up with rather than something we all pretty much accept already. But our everyday set of psychological assumptions is what philosophers call a "folk theory," namely *folk psychology*. These systematically related views about the way internal states are caused by environmental conditions and interact with each other to produce behavior were developed over the generations from the experience of everyday life and handed down to us. Folk psychology was developed as a way of predicting and explaining each others' behavior, and it seems to have served us very well for thousands of years. As Fodor (1987) and Bratman (1987) point out, our everyday lives rely heavily on our folk predictions about what people are likely to believe, desire, and intend in light of various pieces of information about them, and what they are likely to do in light of these internal states.

One thing that may stand out about our folk psychological theory is how much certain of its clauses sound like platitudes or analytic truths. Of course, we can imagine worlds in which there are no intentions, beliefs, desires, feelings of guilt, and so on. But what we cannot imagine is a world in which people have these states but in which they do not have the properties attributed to them by folk psychology. Come to think of it, though, folk psychology is not the only theory which is such that, if its posits do in fact exist, we can know *a priori* that they will have the properties attributed to them by the theory. This seems to be true, for instance, of mass – it would appear that we can know *a priori* that if there really is such a thing as mass, then it causes entities to resist acceleration. Now if quidditists are right about the properties of fundamental physics, it might well be that there could be such a thing as mass that does not resist acceleration. But the difference may seem to be this. Our idea of mass may just be that of whatever it *actually* is that plays the most basic roles that our fundamental physical theories attribute to mass. We are thus guaranteed that if there is such a property it will in fact play the roles by which we pick it out, though the very same property might not have played those roles. On the other hand, our idea of folk psychological states like intentions, beliefs, and desires may just be that of *whatever* plays the most basic roles that our folk psychological theory attributes to them. This would explain why folk psychological states would be so multiply realizable – why an alien being made out of silicon or fluid sacks could have the same psychological states as us, even though his are states of sacks or silicon and ours are states of our neurons.[144]

If this is right, then the way we got to our folk psychological concepts was by having a folk psychological theory couched in terms of those concepts. Building on the work of Ramsey and Carnap, David Lewis (1970, 1972) offered what has become something of a canonical description of how concepts get their meaning from the theories in which they are embedded, which has become known as the Ramsey-Carnap-Lewis [RCL] theory of theoretical concepts.[145] The way we can identify the content of a set of

[144] See for instance Lewis (1980), Fodor (1987), and Braddon-Mitchell and Jackson (2007).
[145] It is usually called the 'RCL theory of theoretical *terms*', but what I care about are the mental states that terms in spoken languages express, rather than anything about those spoken languages. (Of course you might think that concepts are in a sense terms in a syntactically structured language of thought. In fact your simultaneously thinking that and being sympathetic enough to conceptual analysis so as to combine it with

concepts that get their meaning from the theory in which they are embedded is to first take the clauses of that theory and replace the terms that express those concepts in the theory with free variables. Thus, suppose we had a very long list of clauses of our folk psychological theory of the form:

> […., if someone **has a desire** with content *S* (i.e. desires to bring about that *S*) and **has a belief** with the content that *ϕ-ing will bring it about that S*, then *ceteris paribus* she will *ϕ*, …],

we replace the terms expressing our psychological concepts, like 'has a desire', and 'has a belief' with free variables to yield:

> […., if someone is in **state *x*** with content *S* and is in **state *y*** with the content *ϕ-ing will bring it about that S*, then *ceteris paribus* she will *ϕ*, …]

We can then understand our folk psychological theory as claiming that there are states like state x and state y, which is embodied in what Lewis called the Ramsey Sentence:

> There are states *x, y, z,* …such that […., if someone is in **state *x*** with content *S* and is in **state *y*** with the content *ϕ-ing will bring it about that S*, then *ceteris paribus* she will *ϕ*, …]

We can then understand the judgments about beliefs and desires to be judgments that someone is in the states that the Ramsey Sentence claims to exist, thus to think that someone desires something is to think that the content of the Ramsey sentence is true and that the person is in state x with that content, to think that someone believes something is to think that content of the Ramsey sentence is true and person is in state y with that content, and so on.

---

the RCL theory would be something about you that I think I would very much like. It is just that 'terms' has ever so strong an association with spoken language.)

Now not every belief anyone has ever had about someone's beliefs and desires can be counted as parts of folk psychology for this to work. For lots of people have had lots of false beliefs about people's mental states, and this surely does not mean that all of our folk psychological beliefs are false. Nor does it seem at all plausible to count even all law-like regularities as part of the folk theory that gives our folk psychological concepts their content. Freudians used to believe that prepubescent children had subconscious sexual desires and that all adults have incestuous desires, but this did not automatically mean that they had a different notion of desire – surely we token the same notion as they when we deny that children want sex and that most people want to sleep with their parents. Instead, it seems that there is a more specific set of "core roles" that our folk psychology assigns to various mental states which determine the meaning of our concepts of those states, but that by no means exhausts all of the roles that we might think these states play. This is the kind of picture that Braddon-Mitchell (2005) has argued will be true of meaning-giving theories generally, where there is a general meaning-giving theory that can be filled in by more specific theories, which more specific theories can turn out to be false without affecting the meaning of our general theoretical concepts.

If this is the correct account of how our folk psychological concepts get their content, then our moral emotion concepts in particular are the concepts of those states that play the roles assigned to them by the core, meaning-giving portions of our folk psychological theory. Now we saw in the last section that the coordination of our moral emotions and moral judgments would have been ubiquitous features of the emotional and social lives of our ancestors. Indeed much of their adaptive function seems to rely on its being common knowledge that they are ubiquitous in these ways. Flexibly settling on what sorts of defection will be enforcedly deterred requires that each member of a social group draw the entailments about the joint warrant for guilt, anger, and feelings of obligation. But it also seems to require that each member know this about each other member. In order to get on with social life, hominids would have had to be able to expect that their fellows would not do certain things out of feelings of obligation, that they would be apt to feel guilt in certain situations, and that they would tend to anger in others. Intense hominid interaction in large groups requires that the hominids be able to know

that certain kinds of defection will tend not to take place, when and how severely they will be punished when they do, and whether and to what extent this punishment will be accepted or resisted. Considerations of social coordination also suggest that each hominid must know that each other knows this, and that each other knows that he knows that he knows this, and so on.

If common knowledge that people take moral emotions to be warranted in a package was this important to the adaptive function of their normative control, it might well be one of the most important things for a hominid to know about these states. At the same time each hominid would not simply be an outsider, theorizing about the normative beliefs of others, but an intense normative participant, obeying – for reasons we have seen – the constraint of thinking moral emotions warranted in a package himself. For these reasons, one of the most central things a hominid might think about moral emotions was that they were the states that had this package of joint-warrants – that in addition to feeling bad and motivating amends, guilt is that state that's only warranted towards those acts that warrant outrage on the parts of others, and which are such that you should have felt obligated not to perform them in the first place. As central not only to normative thought but to explanation and prediction as well, we might thus expect this kind of clause to have made it into the central, meaning-giving portions of our folk psychological theory of guilt, and to have become part of its content in the way described by the RCL theory of theoretical concepts.

Similar remarks might be made about feelings of obligation as states, rational mandates for which entail the existence of mandates for strongest motivation in their direction. For if this is what was needed to solve the problem of insufficiently reliable deterrence, it would have been important not just for everyone to infer mandates for strongest motivation from mandates for feelings of obligation, but for everyone to know that people would tend to draw them too. As such, we could expect it to have become part of the meaning-giving core of folk psychology that, in addition to having dthat[indicating feeling] phenomenal character and involving motivations to do *X*, feeling obligated to do *X* is the kind of state that is such that if it is rationally mandatory to feel it, it is rationally mandatory to be most strongly motivated to do *X*.

Now of course hominids would have had all kinds of other views about moral emotions and their warrant. They would have tended to think, for instance, that people should and usually will feel obligated not to steal their neighbors' food if they are tempted to do so, and that people should and usually will feel guilt for beating up their elderly parents. But we would not expect these kinds of views to make it into the meaning-giving core of folk psychology. These are exactly the kind of substantive views that are too specific for a general-purpose tool like normative judgment or folk psychology to get tied down to. For just as normative judgments and correlative emotional proclivities need to be applicable to all sorts of contingencies, so too does the folk psychological theory of those judgments and proclivities. We would want our hominids to be able to interpret the thoughts of everyone who so much as might be engaged in the process of coordinating cooperation through judgments about warrant for moral emotions. Some people might not think that they should feel obligated to do various things but not to steal (or not to do what we might consider stealing), and this is a very important thing to know if you are going to be around them. Some people might even have thoughts about guilt but think that they need not feel guilt for beating the elderly. We are talking about hominids here, and they can come up with all kinds of crazy things, including the view that the elderly are cursed or that it is their fault that they have not yet committed suicide and freed up resources or what have you. It is rather important to interpret the thoughts of these people, especially if you are considering letting them near your elderly parents. So while there would be good reasons to let things like the joint warrant of moral emotions and the Contour Thesis into the meaning giving core of folk psychology, there are good reasons not to let in substantive views about when in particular various moral emotions are warranted.

5.4. Attitude Kinds

What I propose, then, is that the joint governance of our moral emotions was a sufficiently important feature for the purposes of both the prediction of behavior and normative thought itself that it became part of the meaning-giving core of our folk theory,

and became embedded into our folk moral emotion concepts as a result. If this is right, then we can see an important thing that our moral emotions have in common. They are emotional states which are such that there are relations of entailment between their warrants. In this way they form a sort of valenced-attitude-kind. The clustering of our moral emotions in an attitude kind gives us a way to cleave off moral concepts from other ethical concepts. Moral concepts are concepts of warrant for moral emotions, where moral emotions are those whose warrants enter into entailment relationships with those like guilt, anger, and feelings of obligation. We still need to reference specific moral emotions to get at the emotional kind, but that is not a problem. These emotions also have a distinctive phenomenology, motivational tendency, and attentional focus; the entailment relations between warrants just allows us to group them together in a kind.

There are reasons to believe that there are other attitude kinds of this sort. Consider what I have been calling "pro-attitudes towards" a state of affairs *S* – desire, wish, hope, and being glad that *S* obtains. Intuitively there are entailments between warrants for these states. It's fitting to desire *S* when *S* is in the offing iff it's fitting to hope for *S* when it might happen iff it's fitting to wish for *S* if it's a distant possibility iff it's fitting to be glad that *S* when *S* obtains. States of desire, wish, hope, and gladness had an adaptive function too, though one that is much older than anything specialized to hominid evolution. They are presumably states that help set things for organisms to try to bring about and keep about when they can. Desire motivates you to bring them about when you actually might be able to, and gladness motivates you to keep them in existence once they are, and wishing and hoping help get you ready for a day when they are in the offing or have obtained. So it would only make adaptive sense for an organism to bear these attitudes towards the same states depending upon whether they are distant possibilities, nearby possibilities, or ways the world actually is. As in the case of the normative governance of moral emotions, the normative governance of these pro-attitudes has the adaptive function of enabling them to play their same goal-setting roles but in a way that is more flexible and better suited to complex and variable environmental conditions. But this will only be possible if they are normatively governed as a set – if, that is, we think that it is fitting to desire *S* when *S* is in the offing iff we also think that it is fitting to be glad that *S* when *S* obtains… That people think this way is moreover a

very important thing to know them if you want to predict and explain their behavior. It pays to be able to infer from Jones's fantasizing about *S* that he may well aim at bringing it about if he can, and from both of these that he is likely to keep *S* in existence once it obtains. We want to be able to interpret the thoughts of all sorts of people with different hopes and desires, but there is no clear reason why we'd need to interpret the thoughts of people whose hopes and desires themselves radically diverge.[146]

One thing worth mentioning is that, even if we do incorporate clauses about shared warrants into our folk concepts of valenced attitudes, nothing prevents us from applying those folk concepts to beings whose mental states we do not hold to standards of warrant. Thus dogs and six month olds will have all kinds of wishes, hopes, desires, and feelings of gladness.[147] Many of us might not, however, want to say in any literal sense that it would be unfitting or irrational for dogs and infants to desire certain states. We can talk playfully of "crazy" dogs and babies, and we can talk seriously of dogs and babies who have unhealthy or dangerous desires (like a desire to ingest plastic or antifreeze). But as important as the latter kind of talk is, it is not literal talk of the fittingness or rationality of their attitudes. (Its import would rather appear to consist in the attitudes it is fitting for us to have – like a desire to get the dog or baby help if we can, and to protect it from the desire if we cannot). Yet nothing in our notion of a state which it is fitting to be in *if* it is fitting to be in all these other states rules our the attribution of desires and hopes to dogs and infants. Since it is neither fitting nor unfitting for these beings to be in *any* of these states, it does not violate any of the fittingness requirements for these beings to be in any one of them. We still pick out states in these beings with the same phenomenology, motivations, and attentional focus as those in us – those that, in us, are unfitting unless the corresponding states are fitting too.

In this way the idea that our folk emotion concepts incorporate views about connections between their warrants has an important advantage over the judgmentalist idea that our emotions involve ethical judgments, and indeed over the quasi-judgmentalist

---

[146] The theoretical demands on folk psychology were perhaps not those of abnormal psychology.

[147] Anthropomorphizing or "Adult-opomorphizing"? Only in the sense that non-solipsism is "me-omorphizing." The idea that spoken language creates a difference here is a joke – how do you know that the noises uttered before you are words in a language that express the mental states you assume they express? If you need more than simulation and explanation of behavior you have neural correlates, which would seem to work as well by way of evidence in dogs and infants as in other (apparently) psychologically typical adult humans.

picture that they involve representations of ethical facts. The former view entails that, since infants and non-human animals lack our ethical concepts, they cannot be in the same emotional states as us. The latter view entails that, unless infants and non-human animals (and adult humans too) somehow represent ethical facts, presumably by entering into nomic or ontically explanatory relationships with them, then they will also fail to be in these emotional states. The view that folk emotion concepts involve clauses about shared warrants, on the other hand, requires neither that a being be capable of having thoughts about ethics or warrant, nor that she otherwise represent ethical facts, in order for her to have the same emotions as us. For a being can be in a state that would be warranted only if another is without having the slightest notion of what warrant would be, and without in any way representing anything about warrant or ethics.

5.5. The Honor System

As I mentioned earlier, we seem to have a set of appraisals of conduct distinct from but in many ways similar to moral appraisals. These include assessments of conduct as lowly, shameful, contemptible, and non-morally virtuous. I think that these kinds of assessments have parallels with moral assessments that are rather striking. In particular, I suspect that these appraisals can be analyzed in terms of the fittingness of attitudes like shame, scorn, motivation by a sense of honor, and non-moral esteem and disesteem. For similar reasons to those given above, I suspect that these attitudes too form an attitude kind similar to but distinct from that formed by the moral emotions. Understanding this may help us to see how these other ethical concepts can be as relevant to what we have reason to do as our moral concepts, yet how they might (as I think we usually assume) pertain to a somewhat different set of practical considerations.

Take first the idea of shamefulness. Now the idea of what is shameful can easily be confused with the idea of what is embarrassing, as well, it might seem, as the idea of something about oneself that is aesthetically displeasing or even downright depressing. Indeed, I suspect that our English term ‘shame’ does not distinguish very sharply between

what might be thought warranted by these kinds of thoughts. In certain contexts, we may use 'shame' to refer to both embarrassment and negative aesthetic self-appraisal (particularly as in 'I'm ashamed of you' and 'I'm ashamed of the way my foot looks'). But I think that we have a distinct concept here, which I want to try to insolate, and, once isolated, I shall stipulate that I shall use 'shame' to refer to it alone.

We seem to have the idea of an intuitively "shame-like" negative attitude towards our actions which is distinct from embarrassment in that it can be felt in private. It is difficult if not impossible to be embarrassed about having done something if one does not somehow represent others as seeing one's doing it. Shame, on the other hand, can easily be felt in private – you might have successfully tricked everyone into thinking that you are very brave, but feel ashamed about how cowardly you really are. Now of course we can also think that certain things about ourselves are aesthetically ugly, and this can be (and in fact usually is) a source of serious negative affect. This is very often how people feel about their physical characteristics like their height, facial features, and sexual characteristics, but also things like the sounds of their voices, their lack of innate intelligence, and their lack of innate musical talent or athletic ability. Feeling bad about being what one regards as ugly or deficient in an important respect can involve feelings of depression, motivation to conceal one's apparently problematic characteristics, and inclination to withdraw oneself from others.

Now I think that in one perfectly good sense of the English term, this kind of negative attitude towards one's characteristics can be called 'shame'. But it is not the sense of the term in which I am interested. That the idea of this kind of negative attitude is not the only one that we can express with 'shame' might be suggested by the awkwardness of parlaying everyday thoughts about shame in this sense into thoughts about what we might call shameful. Many people are ashamed of their height or facial features but the same people would probably find it a bit odd to speak of height or facial features as shameful. Embarrassing, ugly, and depressing, yes, but *shameful*? Nor do I think this is typically just a manifestation of emotional recalcitrance. People who in one perfectly good sense of the term are ashamed of things like their height or facial features may well think that it makes perfectly good sense to feel the way they do. People in our society are rather cagey about admitting to being this "shallow," but we could well

imagine someone (perhaps with a mental life a bit more like that of an ancient Greek) admitting the following:

> "Good height and facial features are important things to have in life; your life is poorer without them, and others won't want to have as much to do with you. And who could blame them? It really is nicer to have people with the right height and facial features for your friends, and especially for your romantic partners."

Yet the very same kind of person might still find it awkward to speak of height and facial features as shameful, or something one should be ashamed of. To bring things perhaps nearer to home, many people in our society would be much more willing to admit things like the above about the lack of innate intelligence or talent. It makes perfect sense, they might say, to be disappointed that one does not have these things, for they are important in life and highly desirable things to have in one's friends or partner. But they would probably find it odd to say that low innate intelligence or lack of innate talent are things that are shameful. They are, again, embarrassing, unfortunate, and perhaps even depressing, but something can seem very strange about calling them shameful.

What this suggests, I think, is that there is an attitude that we can refer to when we speak of 'shame', which is distinct from that which people typically have towards those physical features of themselves that they find deficient. What one will typically be ashamed of in this second sense will be one's actions and those abilities and other features of oneself that are influenced by one's intentional actions. To help fix ideas about these attitudes, we might engage in a bit of preliminary evolutionary speculation. Social hierarchies are important to the lives of all social animals, and social primates are an excellent example. Whether the hierarchies are structured by physical dominance, fertility, or kinship, it paid the genes to program organisms to be able to coordinate on a pecking order. Everyone saves a lot of conflict and expense if the inevitable losers of certain conflicts simply "know their place." Now this seems quite plausible as a story about how 'shame' in the first sense (i.e. in which people are ashamed of their physical characteristics) came to play the role it does in our emotional lives. Knowing you don't have what it takes to win certain competitions for status and mates causes you to have a

sort of slinking, demurring demeanor, rather than the sort of haughty, jaunty, air of those who are in a position to win out over you, and compete with each other for top spots in the hierarchy.

But the kind of shame that we often feel about our actions and those abilities of ours that can be shaped by our actions might seem to play a slightly different role in our lives. For wherever we may be in the social hierarchy, we still face a different sort of challenge, namely the need to overcome inclinations like laziness and fear in the service of important projects. While motivations regarding hierarchy are needed and widely dispersed throughout the animal kingdom, motivations regarding the need for self-control and development are rather unique to hominids with their system of normative control and capacity for highly flexible, long-range planning. Indeed, it seems that the kind of shame we associate particularly strongly with our idea of shamefulness is a kind of negative self-sanction that we feel towards things like failures to control our fear and laziness, and failures to exercise the appropriate discipline so as to develop our abilities. Of course, we can feel shame even in this sense towards all sorts of things, including things that we take to be no such failures. But I think that evolutionary considerations suggest that we have here a conative syndrome distinct from that related to hierarchy, which came into our lives to deal with a different adaptive problem. This kind of evolutionary speculation would help explain how 'shame' in the sense of the attitude historically related to hierarchy and 'shame' in the alternative sense of the attitude that was historically related to self-control and development would have come to have different typical (but by no means "formal") objects.

An important advantage of distinguishing the kind of shame bound up with our judgments about what is shameful from that related to social hierarchy is an ability to make sense of the way most people tend to think and feel about the kinds of things that they take to be the appropriate objects of shame in these two distinct senses. Certain people who are beautiful, intelligent, and talented may in a sense "feel superior to" those who are ugly and innately ill endowed with intelligence and talents. But far fewer – perhaps all who are not cruel children or bullies – would think to treat them with scorn or disdain simply on account of their looks or innate endowments. A nineteenth century

nobleman might feel far superior to his servants in terms of status, but it would be quite another thing for him to actually disdain, scorn, or despise them simply on account of their humble birth. He would be far more likely to reserve these attitudes for commoners who he took to be too ill disciplined to control their drinking, or too lazy to find and keep honest work. In the same way a military officer from the same period might well feel himself superior to his common soldiers, but it would be quite another thing for him to despise his soldiers on account of their commonness. That is something he would be more likely to reserve for those who broke and ran in the heat of battle, something he would be very unlikely to feel towards those who steadfastly stood in line.

One way to understand these very different sorts of responses would be to observe that the extent to which to whom you are relating to is possessed of self-mastery would have been adaptively relevant to an early hominid in a very different way than where that person stood in relation to him in a social hierarchy. For especially if group size is not astronomical, a hominid would have had to engage in cooperative projects with others of at all levels of the social hierarchy. What matters for cooperative projects is not whether people have innate characteristics that mark them as only so fit to win competitions for social status. What matter is rather what one's partners do with the innate gifts that they have. But if a potential cooperation partner shows himself to be too undisciplined to contribute adequately in his role, then one would do best to exclude him from cooperative ventures until he has demonstrated an ability to contribute. Scorn or disdain motivates one to cut ties with someone until he has shown himself to be worthy. In a corresponding way, shame in the sense in which we are interested primarily involves motivations to either display or develop one's abilities in response to a perceived but corrigible deficiency.

Our notion of what is shameful, then, is that of what warrants shame in the sense in which we are interested. But, as we might expect, to think that something is shameful also seems to involve thinking that others are justified in having a response of scorn or disdain to it. If this is right, then we will have a fitting attitude analysis of what is shameful that parallels our previous fitting attitude analysis of what is morally blameworthy:

Fitting Attitude Analysis of Shamefulness:

To judge that someone's act, motivation, or character trait is shameful is to judge that it is fitting for him to feel shame for performing or having it, and fitting for others to feel scorn towards him for performing or having it.[148]

We might thus contrast the way in which our judgments about the fittingness of guilt and anger are adaptive responses to the same kind of behavior – a defection – while judgments about the fittingness of shame and scorn are adaptive responses to a different kind of behavior – a lack of effort or a corrigible disability. Gibbard (1990) drew attention to exactly this in his contrast between guilt as a response to portending punishment and shame as a response to portending exclusion. My main concern is only that I think Gibbard's discussion did not distinguish adequately between the two different senses in which we can feel ashamed, where it is only the second which is concerned primarily with ability to contribute to cooperative schemes. Shame in this second sense is actually tied rather closely to something someone can do, at the very least over the long term, where it is only the first sense of shame that is primarily concerned with having to cope with incorrigible defects and motivate a "willingness to accept poor terms in cooperative schemes when better terms will not be forthcoming."[149] This is, I believe, important for understanding how shame and scorn align with more forward-looking motivations, which are ultimately central to our practical reasoning.

Just as we have an idea of moral wrongness that is distinct from that of moral blameworthiness, I think that we have the idea of something that is *lowly* or *beneath oneself* that is distinct from that of its being shameful. We might tend to think, for example that people with certain mental illnesses or overpowering addictions often do things that are lowly or beneath themselves, but that that it would be unfair to have scorn or disdain for them for doing what they do. Consider for instance an abused person who acquiesces in the most degrading forms of treatment so as not to lose her abusive partner. Or consider a drug addict who could safely quit his addiction, but who, due to an

[148] Gibbard (1990) gives exactly this sort of analysis of shamefulness.
[149] See (Gibbard 1990, 298).

uncontrollable urge for the drugs offers to do the most demeaning things to obtain them. These things, we might think, are lowly or base, but that it would be grossly unfair to look down on their authors as we might be inclined to look down upon a mentally healthy glutton or lazybones. For these poor people are not in some relevantly strong sense fully responsible for what they have done. We might be justified in being wary of what these people will do, but not in abandoning them simply on account of their "deserving" such abandonment, which is the sort of thing that scorn or disdain would seem to motivate. As these acts thus do not befit scorn, they remain lowly, but they cannot be shameful.

What, then, is it to think that something is lowly? As we saw, we can think that certain things are lowly in ways that seem to imply nothing about their moral status, as when we think this about acts of cowardice, weakness, and laziness that do not harm others. Now thinking that an act would be lowly, base, or beneath oneself seems to involve thinking that one should be averse to doing it. This kind of aversion is what people used to describe as "being moved not to do it by one's sense of honor" or even "feeling honor-bound not to do it." The kind of aversion that one thinks it appropriate to have towards performing lowly acts is similar to yet distinct from feeling obligated not to perform them. Feeling obligated not to do something is, as I described it, a kind of prospective guilt-tinged aversion to doing it. Feeling honor bound not to do something, on the other hand, is more like a kind of prospective shame-tinged aversion towards doing it.

Indeed, the same considerations that we have seen in favor of fitting attitude analyses of moral concepts would seem to count in favor of understanding thoughts about lowliness as none other than thoughts about the fittingness of feeling honor-bound not to do certain things. This would explain the way in which thoughts about lowliness guide our sense of honor and our intentions to avoid doing what we take to be lowly. It would also help us explain what we mean when we say that it is lowly but not shameful for the abused person and the drug addict to do what they do. For although it would be unfair of us to feel scorn for them, and they need not feel ashamed given their conditions, it remains true that these people had reason to feel honor-bound not to do what they did, and had corresponding reason not to do it.

Just like we needed to distinguish feeling obligated not to do something from things like fears of punishment and desires to avoid feeling guilt, so too we must distinguish feeling honor bound to not to do something from things like fears of shunning and desires to avoid feeling shame. If you think that an instance of begging someone for help or pleading for mercy is lowly or beneath you, you will have a propensity to have a kind of shame-tinged aversion to doing them, which will motivate you not to do them even if no one will find out about them and even if you knew you could take a pill that would prevent you from feeling shame for doing them after the fact. Like feeling obligated not to do something, the kind of prospective shame-tinged aversion to do something involved in feeling honor-bound not to do it is either intrinsic or to towards doing the thing as a way of doing something else that one feels intrinsically honor bound not to do. Thus, if something is intrinsically lowly, understood as an act that it is fitting to feel intrinsically honor bound not to do, we have by the Warrant Composition and Rational Ends Principles that it is worth avoiding doing that thing as an end in itself.

Now as with the fittingness of feeling obligated not to do something, it seems that we can distinguish between two senses in which it can be fitting to feel honor bound not to do it. Suppose, for instance, that very bad things will happen to you personally unless you degradingly beg someone for help or plead someone for mercy. It seems coherent to think that in such a situation you would be justified in feeling honor bound not to beg or plead, but that, given the stakes, you would also be justified in having no such feeling. On the one hand this is a deeply degrading an inelegant thing to do, and one would be doing it only to save one's own hide, so a person might well be moved by a sense of honor not to do them. But on the other hand, we might have someone who would feel honor bound to avoid doing these sorts of things in lower stakes situations, but who does not feel that sacrificing everything simply so as to avoid begging or pleading is something he has to do. It seems, however, that we can coherent accept that both kinds of people would be justified in their feelings. We might think that the fact that one would be degradingly and inelegantly begging or pleading justifies feeling honor bound not to do it, but given the fact that the alternative is, say, death, we might think that these considerations are insufficient to make it mandatory to feel honor bound not to do the things in question.

Ordinary English might not distinguish too sharply between different concepts that concern warrant for feeling honor bound to do things. But I think that we do have a concept, which we might call that of a LOWLY ACT, a BASE ACT, or a LOW-DOWN THING TO DO which we can analyze as that of something that that one must feel honor bound not to do. Or, to be more precise, it is that of an act that one must feel honor bound not to do if one is not already going to omit doing it. It might be lowly for me to grovel for certain favors from people, but it would never even enter my mind to do so, in which case it would seem that I am not flouting any rational mandates about what I should feel honor bound not to do. So, very much like the way in which we analyzed the notion of a morally wrongful act as that of an act that it is rationally mandatory to feel obligated to omit unless one is already going to omit it, it seems that we can analyze the notion of a lowly act as one that it is similarly mandatory to feel honor bound to omit:

> **Fitting Attitude Analysis of Lowliness**:
> To judge that agent *A*'s act of ϕ-ing is lowly is to judge that, unless *A* is already going to refrain from ϕ-ing anyway, it is rationally mandatory for *A* to feel honor bound not to ϕ (or equivalently: to judge that, unless *A* is already going to refrain from ϕ-ing, it is rationally mandatory for *A* to feel prospective shame-tinged aversion towards ϕ-ing).

Now as our above observations might suggest, rational mandates for being honor bound seem, like rational mandates for feeling obligated and unlike rational mandates for desires for states of affairs, to be sensitive to consideration that weigh in favor of doing other things. That is, the fact that one must grovel to save one's life seems to count, not just in favor of being most strongly motivated to grovel, but against having to feel honor bound not to. If this is right, then considerations that contribute to having to feel honor bound to do something will only succeed in making it the case that one must feel this way if they are strong enough to outweigh reasons to be motivated to the contrary. If this is right, then just like we had a Contour Thesis asserting the dependence of rational mandates for feelings of obligation on rational mandates for strongest motivation in their

direction, we will have the following kind of contour thesis governing rational mandates for feeling honor bound:

> **Parallel Contour Thesis**:
> If it is rationally mandatory for agent *A* to feel honor bound to do *X*, then it is rationally mandatory for *A* to be most strongly motivated to do *X*.

In a way exactly parallel to our demonstration of conclusive reasons not to do what is morally wrong in chapter 4, the fitting attitude analysis of lowliness, Parallel Contour Thesis, Most-Motivation-Action Principle, and Motivation Partition Principle together entail that if an act is lowly, then one has conclusive reason not to perform it.

If this is right, then the fact that an action is lowly or base has the exact same kind of practical import as the fact that it is morally wrongful. Too often I think that there is a tendency to look down on considerations of what would be a lowly thing – a cowardly, weak, or indulgent thing – as something like the poor cousins of the family of ethical considerations. At best, people seem to think that these kinds of considerations have to be shown to relate to something else for them to constitute reasons not to do something. Unless it can somehow be shown that doing what is lowly is worse for the actor, or worse for everyone, or in any event flouting a consideration of the kind that could make an act morally wrong, then whatever makes it lowly can be no genuine reason not to do it. At worst, people seem to think that considerations of what is lowly or beneath one are obsessions with reputation or bizarre aesthetic ideals of what kinds of things make one have an "ugly" character or something like that.

But if I am correct, considerations of what is lowly or base need no independent support by other considerations in order to constitute reasons for action. Just as with the considerations that make something intrinsically morally wrong, the considerations that make something intrinsically lowly – that make it fitting to feel intrinsically honor bound not to do it – will be reasons to avoid doing it for its own sake. Whether something is lowly is not just an aesthetic assessment or a concern about reputation; it is a concern about what serves a rational end in itself, an omission of something that it is fitting to be intrinsically motivated not to do by one's sense of honor. These assessments are

moreover sensitive to one's overall set of reasons for action. Rather than rigid aesthetic ideals of avoiding what somehow feels dirty to do, they are realistic assessments of whether, given the various concerns we should have, considerations such as those of what would be acquiescence to weakness or being demeaned are strong enough to tip the balance against doing something. I see no reason to doubt our intuitions that these kinds of features can make actions lowly, and thus to be avoided. Such considerations are thus independent source of practical concern, though they are surely not the only ones, and can often be overridden. It is simply that when this happens our action cannot be criticized as lowly, for what else were we to do?

I think that there are other ethical concepts that link up to those of shamefulness and lowliness in ways parallel to those in which moral goodness, badness, and supererogation link up to moral blameworthiness and wrongness. First, I think that we have a notion of what is non-morally virtuous or estimable. This, as I mentioned in chapter 1, is how we tend to feel towards impressive exercises of personal courage, and intellectual and athletic accomplishments, which may very well have been in no way aimed at a moral cause or making the world a better place. Correspondingly, I think that we have a notion of what is non-morally vicious or contemptible. This is how we might feel towards certain instances of weakness or ill discipline. The general considerations we have adduced in favor of fitting attitude analyses would seem to support the following analyses of these notions:

Fitting Attitude Analysis of Non-Moral Virtue / Estimability:
Let Φ be an action motivated in a particular way, a motive, a character trait, or an agent. To judge that Φ is non-morally virtuous or estimable is to judge that it is fitting to non-morally esteem Φ.

Fitting Attitude Analysis of Non-Moral Vice / Contemptibility:
Let Φ be an action motivated in a particular way, a motive, a character trait, or an agent. To judge that Φ is non-morally vicious or contemptible is to judge that it is fitting to have contempt or non-moral disesteem for Φ.

Like morally esteeming something, non-morally esteeming it involves a feeling with a phenomenal character that we might describe as "looking up to" "feeling impressed by" or "standing in awe" of it. It also involves tendencies to wishfully imagine or fantasize about doing or being like the object of one's non-moral esteem, and involves motivation to emulate it. But when one feels non-moral esteem towards someone, one feels more like congratulating the person whose act, motive, trait, or self is its object for doing or being as she is. In this sense it lacks the peculiar moralistic tincture associated with feeling like thanking her. Of course, one can feel both kinds of esteem towards the very same object, say a person who has just pulled off an especially risky and technically difficult rescue mission. These kinds of esteem remain distinct, however, as we can see in the contrast between having one but not the other towards something. We can easily imagine having non-moral esteem towards people who accomplish great athletic and intellectual feats as a result of a significant hard work and discipline, but where they were not responding in any way to moral considerations or considerations about someone's welfare (so, for instance, none of this "Gipper" or dying wish of the coach stuff). Similarly, we can imagine having moral esteem towards people who do good for others at great personal cost, but where what they did was actually quite easy and took no discipline or self-control at all. Suppose, for instance, that Freddy and a child he does not know are both tied to branches of trolley tracks. The trolley is headed towards the child, but all Freddy has to do to save her is to press a button by his side, diverting the trolley over himself. Suppose that Freddy is just a naturally sympathetic person, who never really had to work on mastering any selfish impulses, and he is so taken with care and concern for the child that he just naturally and without any hesitation presses the button and gets himself run over in the child's stead. We can, I think, quite naturally imagine having a great deal of moral esteem towards Freddy, involving something resembling feeling like thanking him. But we might well imagine having no kind of non-moral esteem towards him; no feeling like congratulating him for pulling off any sort of difficult feat.[150]

[150] I shall, as always, decline to speculate as to whether someone like Kant might have been inclined to doubt the moral estimability of people like Freddy as, among other things, a result of conflating moral

Similarly, contempt or non-moral disesteem is similar to moral disesteem in that it has a phenomenology that seems characterizable as "looking down upon" its object, it involves tendencies to fantasize or wishfully imagine oneself unlike its object, and it involves motivations to disemulate its object. But whereas morally disesteeming someone might be characterizable as feeling as though that person is a "jerk", having contempt for someone involves feeling more like he is "soft" or "a wimp." Of course, just as we can morally and non-morally esteem the same thing, we can have moral esteem and contempt for the same thing too. Using the money one was going to use to pay for an impoverished employee's medical expenses on drugs might be a prime candidate. But our ability to feel them towards different things helps illustrate how they are distinct.

It should be clarified that contemptibility seems distinct from shamefulness and lowliness in at least two important respects. First, certain kinds of acts can be so very weak or cowardly that, while they are certainly lowly, they may be quite literally beneath contempt. For something to be contemptible it must warrant disesteem, and some things may be so low down that it is not even fitting to wishfully picture oneself not doing them – because of course one would never dream of doing them in the first place! Second, for reasons similar to those we saw above, certain lowly acts might fail to merit not only scorn but also contempt. This might be true of someone so deep into an addiction that we realize that there is really nothing he could have done to resist urges to do lowly things. In such a case we might hope that we would be able to resist these urges were we in his position, but we might have to understand that we probably could not. It might take genuinely heroic effort to avoid doing lowly things in his position, in which case we might have no business feeling as though this person is weak or soft for doing them.

Finally, just as we have a notion of supererogation or going above and beyond the call of duty, I think that we have a notion of going above and beyond the call of honor. Although in contemporary English 'noble' seems often to have a moralistic flavor, try saying 'noble' while thinking about the way aristocrats used to fancy themselves and without thinking about self-sacrifice. Or try thinking about what in Greek they used to

---

estimaiblity with non-moral estimability. But if someone were tempted to do this, we might well be understand how she would. What she would get right (if my intuition serves me correctly) is the fact that Freddy is in one sense not estimable. But what she would get wrong (again if my intuition serves correctly) is the fact that this does not mean that there is no sense in which he is estimable. In fact, Freddy is estimable in the sense that is more intimately (in fact inferentially) associated with morality.

call *kalon* or fine action. We might then call an act that goes above and beyond the requirements of honor a NOBLE or FINE act. To get some idea of what this might look like, consider cases in which it might seem plausible to think that it is fitting but not rationally mandatory to feel honor bound to do something (which is not itself lowly). We might consider in this respect the behavior of the fictional character Kunta Kinte in the movie *Roots*. In the movie a slaver tries to whip Kunta Kinte to submit to answer to the name 'Toby' rather than his own name. Kunta Kinte refuses to be broken for as long as he can, knowing that this will cause him to be whipped further. Now one might, in such a case, think that it would not have been lowly for Kunta Kinte to answer to 'Toby' immediately, since he really has little hope of avoiding having to do so eventually. But one might think that Kunta Kinte's refusal to submit was still well supported by reasons – reasons of the kind that could have made submission lowly had the stakes been lower or had eventual submission not been inevitable. In this respect, we might think, Kunta Kinte did what was favored by the kinds of reasons he would have not to do something lowly. He just did what was favored by these reasons even more strongly than another non-lowly option, namely that of saying 'Toby' immediately.

Much as with supererogatory action and good action, noble action may be quite distinct from non-morally virtuous action, in that sometimes just doing the non-lowly thing for the right reasons is virtuous. This might well be so if one resisted one's drug addiction and refrained from doing demeaning things to get drugs. It also seems possible to engage in noble actions in ways that are not virtuous. In *Roots* Kunta Kinte's action is (if my intuitions serve correctly) done for the reasons that make it noble and thus it is a virtuous action. But we might imagine an alternate version of the story in which someone in Kunta Kinte's position refused to submit only because he irrationally believed that saying 'Toby' would curse him, and he fears the curse more than the lash. In such a case, we might say that the person still did what was supported by reasons of honor to a greater extent than honor required, even though he did not do it for those reasons or in a way that was virtuous.

Considerations similar to those we adduced in discussing supererogatory action would thus seem to support understanding noble action, or action above and beyond the call of honor in the following way:

Fitting Attitude Analysis of Noble Action:

To judge that it is fine or noble for agent *A* to do *X* is to judge that:

(1) It is honor-system optional for *A* to do *X*

(That is, it is neither lowly for *A* to do *X* nor lowly for *A* not to do *X*. That is, it is neither rationally mandatory for *A* to feel honor bound *not* to do *X* if *A* insufficiently motivated to *not* do *X*, nor rationally mandatory for *A* to feel honor bound *to* do *X* if *A* is insufficiently motivated *to* do *X*)

(2) It is rationally optional for *A* to feel honor bound to do *X*, and

(3) There is another honor-system optional act, *Y*, such that *A* has strictly more reason to feel honor bound to do *X* than *A* has reason to feel honor bound to do *Y*.

Thus honor, like morality, only goes so far by way of demands, or acts that it is rationally mandatory to feel honor bound to perform and thus (by the Parallel Contour Thesis, Most-Motivation-Action Principle, and Motivation Partition Principle) most reason to perform. After that, honor tries to go further, but only by way of recommendations. These are acts which one has strictly stronger reasons of honor to perform – strictly stronger reasons of the kind that count in favor of feeling honor bound to do something – but not necessary more reason on the whole to perform.

If the foregoing is correct, then we would seem to have concepts of what is shameful, lowly, and noble running in parallel to our concepts of what is blameworthy, wrong, and supererogatory. Just as we tend to think of the latter as the categories of 'morality', we might refer to the former as those that comprise 'matters of honor' or 'the honor system'.[151] Now as I mentioned, it seems that the emotions that have figured into our analyses of the concepts of the honor system form an attitude kind among themselves that exists in parallel to that attitude kind constituted by our moral emotions. We have

[151] In my experience contemporary American English speakers (or at least those at a sufficient remove from military circles) often tend to freak out if you talk about 'honor'. The term seems to have gotten closely associated with phrases like 'honor killings', which would stand to what I am calling the honor system as massacring people on the grounds that their ancestors allegedly (and on rather crazy grounds) did culpable wrong would stand to 'morality killings'. I hope that the phrase 'honor system' goes some way to mitigate the freaking out by reminding us of a relatively friendly situation in which people are expected by dint of their own ethical views not to engage in deception that might not only be wrongful but also lowly.

seen that an act can, on account of an honor system analogue of exculpation, be lowly without being shameful. But it seems that, at least when it comes to acts, an act can only be shameful if it is also lowly – it can only warrant shame and scorn after the fact if its author should have felt honor bound not to perform it before the fact. Similarly, as I suggested, if a noble act is performed for the reasons that make it noble, it will be non-morally estimable or virtuous. And, it seems, if an act is contemptible then there will be some reason to feel honor bound not to perform it.

I think that there are reasons to suspect that our honor system emotions came to form an attitude kind in much the way I suggested that our moral emotions came to from an attitude kind. Recall our discussion of the adaptive functions of scorn and shame in our sense. These would have been to respond to failures to control oneself in the service of pursuing important projects. Since shame and scorn would have been adaptive first-person and third-person responses to this same kind of failure, they would have succeeded in playing their adaptive role only if people tended to feel shame towards just those things towards which others felt scorn. Along these lines, the adaptive function of feeling honor bound to prevent these failures of control in the first place – to rally motivation against one's inclinations to stray from important ends when that inclination was lacking. Now when it comes to acts, it would have been adaptive to feel shame about having done something or scorn for someone for having done it only if it was the kind of thing that she should have tried to omit in the first place out a sense of honor. When our judgments of warrant for shame, scorn, and feeling honor bound came on the scene, their adaptive function would have been to enable these attitudes to play their same adaptive roles but in a way that was more flexible and better suited to complex and varied environments. This they would have succeeded in doing only if hominids were such that they judged each of scorn and shame fitting only when the other was, and that it was fitting to feel shame and scorn after the fact of an action only if it was fitting for the actor to feel honor bound to omit it in the first place.

Having these views about the conative syndromes that constitute shame, scorn, and feeling honor bound to do something would have been important not only to normative thinking but to the prediction of each others' behavior as well. For judgments of warrant for these attitudes to successfully facilitate the deterrence and mending of

tendencies to spin out of control, it would have had to be common knowledge that others will think they should scorn just what they think people should be ashamed of, which would also be acts that they thought one should feel honor bound to omit in the first place. As such, there are reasons to think that views about the joint warrant of shame, scorn, and feelings of being honor bound entered into the central, meaning-giving clauses of our folk psychological theory, and that as a result our ordinary concepts of shame, scorn, and feeling honor bound are those of conative syndromes that are jointly warranted in just these ways. Similar remarks could, I think, be made about why the Parallel Contour Thesis is a conceptual truth about our notion of feeling honor bound, in the same way that the Contour Thesis is a conceptual truth about our notion of feeling obligated.

We could then say that an honor system concept is a concept of warrant for an honor system emotion, where an honor system emotion is one whose warrant enters into relations of entailment with emotions like shame, scorn, and feeling honor bound. What this would mean, then, is that we have two families of ethical concepts – morality and the honor system – each of which have duals which are virtually identical in terms of their connection to what we have reason to do. A natural question to ask, then, is why we should care about the distinction between morality and the honor system. One reason I think it is important is that we do make honor system evaluations all the time, but we tend not to recognize their independent bearing on what to do in the way we tend to recognize those of moral assessments. This, I think, not only has the potential to make us insensitive to genuine reasons, I think it also has the potential to make for confused substantive theorizing that tries inappropriately to account for the practical import of the honor system in terms of that of morality or something else.

But if we were to recognize the bearing of both morality and the honor system on the question of what to do, and if we were to take the same things to be required and recommended by both systems, I think that there would be no harm in having one set of concepts rather than two. Perhaps this, or something very much like it, is what some cultures actually have.

5.6. Mixing it up

Recently, certain critics of analyses of ethical concepts in terms of the fittingness of moral emotions have alleged that certain beings who do not have moral emotions make moral judgments.[152] There are two main groups of people who might be thought to fit this description: children who are too young to experience guilt, and people from foreign cultures who may not have this emotion either. Now one thing to notice from the beginning is that these criticisms have focused on guilt in particular rather than anger or feelings of obligation. Anger seems to be widely shared (though we might wonder if it has the right features to be the kind of anger whose warrant would entail warrant for guilt), and no one seems to have been thinking about feelings of obligation. If feelings of obligation are what are really central to our practical reasoning and our reasons to be moral, then it is that attitude in which we should probably be most interested, and it is unclear if the same empirical worries apply to how widely it is shared. But let us assume, for the moment, that they do – that we have suspicions as to whether sufficiently young children and people from foreign cultures are capable of such a thing as feelings of obligation in addition to guilt (and perhaps outrage).

I actually do not think that the example of young children is worth taking very seriously. The alleged evidence that young children make moral judgments consists primarily in the fact that they use moral language and that they distinguish what they call moral from what is "conventional."[153] But the fact that young children use an adult word to refer to something is very poor evidence that they mean exactly what we do. It is common, for instance, for children to use terms like 'dog' to refer to many four-legged creatures that are not dogs. They also distinguish four-legged creatures from other random things, like toasters. Does this threaten an analysis of 'dog' according to which it does not refer to other four-legged creatures?

---

[152] See for instance Nichols (2004).

[153] Incidentally, as Gibbard (2006) points out, the alleged "moral-conventional" distinction appears not to have been very well thought through. Its distinguishing mark seems to be that moral rules cannot be revoked by authority figures. But, of course, there are moral rules that are not like this. If a legitimate authority sets certain rules, it can be morally wrongful to violate them, but morally permissible to do what would violate them if they are revoked by the authority. In the same way, if I set rules for what to do with my property, it may be morally wrongful for you to do what violates them so long as I say so, but morally permissible for you do what would violate them if I give you permission.

Rather, on the hypothesis that moral concepts are concepts of warrant for moral emotions, we would expect children to have heard us use moral language, and to ape our usage. This may be similar to the way in which I ape the usage of tree experts who talk about 'Elms', and Mary, the scientist who lives her whole life in a black and white room, apes the usage of people who talk about 'red'.[154] One thing the children will notice is that we seem to use 'morally wrong' to mean something other than 'in violation of a convention'. Another thing they will notice is that we try to avoid doing what we take to be morally wrong, we encourage people not to do what we say is morally wrong, and we tend to punish people for responsibly doing what we take to be morally wrong (including very young people like themselves, since (a) it often correlates with what we think is undesirable and (b) this is important for later life). Since one usually co-refers by deferring to experts, and children want to be like adults, they will use 'morally wrong' to mean something much like 'whatever it is those adults are calling 'morally wrong'.' But none of this threatens the analysis of moral concepts in terms of warrant for moral emotions. That is how we, the experts use the term. The non-experts can be safely left to fend for themselves. Maybe they get to be said to be talking about the same thing by courtesy. Maybe they get only the metalinguistic proto-concept whatever has the property picked out by ADULT UTTERANCES OF 'MORALLY WRONG'. Either way, their usage can be safely said to be parasitic on that of someone who uses the language to express concepts of warrant for moral emotions.

And even if sharing language with children somehow did force us to abandon this view of what our moral language really means, the role of concepts of warrant for moral emotions in practical reasoning and guidance would still make them play all the normative roles that we might have thought were played by the concepts expressed by our moral language. If so, then when it comes to thinking about what to do, moral language and whatever it does express can go to the devil. You will then forgive philosophers and people seriously concerned about what to do if they start using some kind of language to express these practically vital concepts. It might even consist of homonyms of what our current moral vocabulary. And why exactly, then, is anyone supposed to waste their breath using your original homonyms, except to keep children in

[154] These examples trace to Putnam (1975) and Jackson (1982).

line and to train them up to eventually think in terms of what the alternative homonyms express?

What is far more interesting is what to say about psychologically typical adults from cultures that may not actually have our moral emotions. For instance, Fessler and Haley (2003) suggest that guilt may be absent from certain cultures, like that of a fishing village in Sumatra. It seems, however, that something like shame in the sense we have been discussing and something like outrage do appear to be universal. Let us refer to what they have as shame* and outrage*. Some cultures, apparently, may pair outrage* with shame* much as we pair anger with guilt or scorn with shame. Now if the fitting attitude analyses of ethical concepts that I have been defending are correct, and Sumatrans lack guilt and pair outrage* with shame*, then they cannot have quite the moral or honor system concepts that we have.[155] Indeed, if what I have said about how views about interdependent warrants enter into the meaning-giving clauses of folk psychology to determine our common sense emotion concepts, Sumatrans may not even be able to share with us the notions of the states that they do have, like shame* and outrage*.[156] Moreover, whether the prospective aversion that tends to deter them from doing what they take to warrant shame* and outrage* is more like feeling obligated not to do it or more like feeling honor bound not to do it, they will presumably be unable to think of under the concept of FEELING OBLIGATED or FEELING HONOR BOUND. For that would require that they think that it is the that had to be warranted before the fact for either guilt and outrage or shame and scorn to be warranted. But we may assume that the Sumatrans think that neither guilt and outrage nor shame and scorn jointly depend on something else for their warrants, because they analytically pair warrants for shame* and outrage* instead of warrants for these other combinations.

So is it a problem if Sumatrans don't have our same ethical concepts like MORAL WRONGNESS and LOWLINESS? It would seem to be a problem if we had them coming out

[155] Assuming, for instance, that to have the concept of warrant for guilt you have to have the concept of guilt. This is something that might actually be challenged; judgments about warrant for an attitude might simply have to be states that guide the attitude in the appropriate way, and this might be possible without one's having a concept of that attitude. But it might be awfully surprising if for purposes of third person prediction we did not have the notion of the response that is guided by the judgment of its warrant.
[156] This would create a problem for the kind of solution offered by Gibbard (2006) in terms of near-moral concepts with fixed emotion concepts, like OUTRAGEOUS.

looking like that had no kind of ethical thought whatsoever, because this clearly does not seem to be the case. They certainly do seem to take certain valenced attitudes to be warranted and to feel and act in light of this. It might also be a problem if we couldn't have ethical thoughts that could be in agreement or disagreement with the ethical thinking of the Sumatrans. Beating someone just for fun is wrong and groveling for unimportant favors is lowly, and we would hope that we can find a way to communicate something like this to the Sumatrans and moreover be able to find agreement with them. We would at the very least want to communicate and find agreement with the practical import of these statements – that there are conclusive reasons not to do these sorts of things, and that you do not have these reasons simply because, say, not doing these them will spare you suffering in the long run. We would hope that we could communicate something about the way in which the practical import of our claims is distinctively ethical. Either that you have reason not to do these wrongful or lowly things as ends in themselves, or you have reason not to do them because they are a way of doing something else that is intrinsically wrongful or lowly.

Now if what I have been saying in this chapter about the parallels between morality and the honor system is correct (and the Sumatrans are as hypothesized), it would appear that Sumatrans are having ethical thoughts that are strikingly similar to our own. They have a notion of what warrants shame* and outrage*, which is very much like both our notion of blameworthiness and our notion of shamefulness. We might call this their notion of CULPABILITY*. Since it is shame* that the Sumatrans seem to have, let us assume that the prospective feeling of aversion they have to doing what they take to warrant shame* and outrage* is tinged with shame* and thus rather like feeling honor bound not to do it, and let us call it 'feeling honor bound* not to do it'. There is every reason to believe that Sumatran thoughts about the fittingness of feeling honor bound* will work just like our thoughts about the fittingness of feeling obligated or honor bound. They will have something like the idea of its being rationally mandatory to feel honor bound* not to do something unless you are going to omit it already, which we can call their notion of WICKEDNESS*. Just like feeling honor bound or feeling obligated, one will feel honor bound* either intrinsically or as a way of doing something else that one feels honor bound* to do. It will follow from this by the Warrant Composition and Rational

Ends Principles that we have reason not to do what is intrinsically wicked* as end in itself, and that we have reason not to do what is non-intrinsically wicked* simply as a means to the end of avoiding doing the intrinsically wicked* thing that it is a way of doing. There will be some variety of contour thesis, Contour Thesis*, asserting that it can only be rationally mandatory to feel honor bound* to do something if it is rationally mandatory to be most strongly motivated to do it, from which it will follow by the Most-Motivation-Action and Motivation Partition Principles that there is conclusive reason not to do what is wicked*. And the Sumatrans will think that an act can only be culpable* if it is wicked*, but that there is such a thing as excused wickedness*, so not all acts that are wicked* will be culpable*.

We could go on to do this for the Sumatran duals of other ethical concepts like supererogation / nobility, but one gets the idea. Sumatran ethical thought thus seems to be identical in terms of its practical import to our thinking about morality and the honor system. We also seem very close to being able to share ethical thoughts with them. We are for all practical purposes agreed with the Sumatrans about what is "ethically out" if we can agree on whether we should feel either obligated, or honor bound, or honor bound* not to perform it. In the same way, we can find practical agreement on what is "ethically condemnable" if we can agree on whether we should feel either guilt or shame or shame* if we have done it, and others are justified in feeling outrage or scorn or outrage* at us for doing it. And so on. We can in this way communicate with the Sumatrans by means of these kinds of generalized ethical concepts calling for some kind of negatively valenced, deterrent response (that obeys a contour thesis) before the fact and, barring exculpation, calling for some kind of negatively valenced reparatory response on the part of the actor and a negatively valenced sanctioning response on the part of others.

If we could agree on this much, we might be able to agree even further – namely that each party should stick with the particular deterrent, reparatory, or sanctioning response that she brings to the table. For what business would we have thinking it rationally mandatory for people to have attitudes that they cannot entirely grasp, let alone reason their way to having? We could then continue to use our previous ethical and folk emotion concepts pretty much intact, so long as we took their domains of application to

be restricted to those who had them, and complemented these with parallel thoughts about what should be felt by people with slightly different ethical concepts. Thus, our notion of wrongness as something it's rationally mandatory to feel obligated not to do could remain as what it's mandatory to feel obligated to not do if you are a westerner. Similarly, their notion of wickedness* as something it's rationally mandatory to feel honor bound* not to do could remain as what it's mandatory to feel honor bound* not to do if you are a Sumatran. Our umbrella notion of "ethically out" would then just be the conjunction of what is wrong-or-lowly and wicked* - what you should either feel obligated or honor bound not to do if you are a westerner, and what you should feel honor bound* not to do if you are a Sumatran. That an act is ethically out in this sense would then be a necessary condition for its being ethically condemnable in the sense of befitting of guilt-or-shame by westerners who do it, befitting of shame* by Sumatrans who do it, befitting of outrage-or-scorn on the part of westerners who did not do it, and befitting of outrage* on the part of Sumatrans who did not do it.

Thus, in talking with Sumatrans about what is ethically out or ethically condemnable, we would essentially be talking disjunctively about what is either wrong or lowly and what is either blameworthy or shameful. Assuming that we can arrive at the same basic conclusions about warrant for these kinds of responses, we will arrive at all the ethical agreement we could ever need or want. I see no reason why the prospects for arriving at ethical agreement of this kind would be any dimmer than those for the westerners or Sumatrans arriving at ethical agreement among themselves.

Chapter 6

The Norm Descriptivist Theory of Reasons

So far we have seen various advantages of analyses of ethical concepts in terms of the fittingness of valenced attitudes [FA-analyses]. In conjunction with an understanding of judgments about fittingness as a basic kind of warrant assessment, they can explain the semantic, causal, and epistemic features of our ethical judgments. Mover, in conjunction with the connection between fitting motives and reasons for action developed in chapter 3, these analyses can explain the connection between ethics and reasons for action. But apart from characterizing them by example and by their characteristic functions of direct attitude guidance, we have yet to say anything so far about what more precisely it is to judge that a valenced attitude is fitting.

In a sense this is an advantage, because the considerations we have adduced so far should be attractive to a theorist whatever her views are about the basic semantics and metaphysics of normative concepts and normative facts. It seems to be a desideratum on any theory of ethical judgments and any theory of judgments about fitting attitudes that it be at least compatible with both kinds of judgments manifesting the semantic, epistemic, and causal properties we have surveyed. For this reason, I think that one should recognize the ability of FA-analyses to subsume these features of the former under these features of the latter as an attractive theoretical virtue whatever one's views about what it is to judge that an attitude is fitting or warranted. Even if one thought that the notion of a fitting or warranted attitude resisted all further explanation or analysis, and that it is simply a brute fact that fittingness judgments have the their semantic, epistemic, and causal properties, it would still be good to keep the stock of brute facts as small as possible. This can be achieved by explaining the possession of these properties by ethical

judgments in terms of their reducibility to judgments about the fittingness of attitudes like desires and emotions.

But in another sense silence about what fittingness assessments are can get us into trouble. As attractive as they may be, FA-analyses face an important problem, which has been aptly dubbed 'the wrong kind of reasons [WKR] problem' by Rabinowicz and Ronnow-Rasmussen (2004). We have throughout been appealing to the distinction between the fittingness of an attitude and pragmatic reasons to have it. Thus, to recall, consider a situation in which an evil demon were to threaten to harm your loved ones unless he detects that you desire that you have an even rather than odd number of hairs on your head. The fact that the demon will harm your loved ones if you do not desire a state in which you have an even number of hairs gives you pragmatic reason to have the desire but does not contribute to its fittingness. There is, however, a worry that what distinguishes those reasons to have an attitude that contribute to its fittingness from those that do not itself involves the concepts the FA-analyst is trying to analyze. Why do one's reasons to, say, desire states in which puppies are happy contribute to the fittingness of such desires, but one's reasons to desire states in which one has an even number of hairs in response to demonic threats do nothing of the kind? A natural explanation might *just be* that the former states are good and the latter states are not, and that what distinguishes fittingness from non-fittingness reasons for desires is that the former are the reasons one has to desire *good* states of affairs. But if this is what the FA-analyst must say to distinguish fittingness from non-fittingness reasons, she will run in a vicious circle when it comes time to explain which kinds of reasons for attitudes she is talking about in her analyses. The problem for the FA-analyst of explaining what distinguishes fittingness from non-fittingness reasons to have attitudes, without running into the vicious circularity of invoking the ethical concepts she is trying to analyze, is the WKR problem.

In the earlier chapters I appealed to a propensity that fittingness assessments have to directly guide our attitudes, which judgments about pragmatic reasons lack, in order to help distinguish the former from the latter. That is, judging an attitude fitting can cause you to have it without your having to do anything to bring this about, where merely judging that you have pragmatic reason to have an attitude characteristically cannot have this effect. But one might well think that this cannot be the whole story. For it might

seem to be a rather contingent feature about our psychology that our judgments about pragmatic reasons for attitudes are unable to have certain direct effects on our coming to have them. Moreover, we have seen in chapter 4 a way in which certain assessments of attitudes, which we might call a species of judgment about "reasons to have them" can typically cause us to have the attitudes in question without our having to do anything to bring this about. This was so of judging an attitude estimable or disestimable. We might say, following D'Arms and Jacobson (2000) that these are something like "moral reasons" to have – or to be such as to have – the attitudes in question (which must be sharply distinguished, of course, from the moral reasons that count in favor of doing things by contributing to the actual fittingness of moral emotions that involve motivations to do them).

But if the bare fact of whether a judgment requires action to instill an attitude is insufficient to distinguish judgments about fittingness from judgments about other kinds of reasons, how could else could we explain the distinction without invoking the ethical concepts that the FA analyst is trying to explain in terms of fittingness assessments? In this chapter I will begin by reviewing some of the attempts that have been made to explain the difference between judgments about reasons that contribute to an attitude's fittingness and other kinds of reasons in terms of the content of those reasons, or the kind of considerations they are. We shall see reasons to think that such attempts are ultimately unsuccessful; and so much so, in fact, that some have thought that the prospects for non-circular FA-analyses are dim. But I shall argue that closer attention to the role of fittingness assessments in directly guiding our attitudes provides an alternative way to understand what it really is to judge an attitude fitting as opposed to supported by other considerations. I shall argue that there are reasons to analyze fittingness assessments in terms of a particular kind of functional state, namely the acceptance of norms for attitudes. While things other than norms for an attitude can in principle (and sometimes in practice) give rise to it without the intervention of motivated behavior, the direct governance of our attitudes is part of the role of accepting a norm, and the appropriate operation of such a system of governance is both necessary and sufficient for a judgment that an attitude is fitting.

I will conclude by arguing in favor of a particular way of understanding fittingness assessments – and indeed all normative assessments – in terms of the acceptance of norms. According to this view, which I call 'Norm Descriptivism', judgments that it is fitting for an agent to have an attitude are judgments that the attitude is prescribed by the most fundamental norms that that agent deeply accepts. My contention will be that Norm Descriptivism offers us the best explanation of how fittingness judgments guide attitudes, what we are doing when we engage in basic inquiry into which attitudes are fitting, and how such inquiry can hook onto facts about fittingness. I will also argue that Norm Descriptivism best explains what is distinctive about attributing reasons to agents and which entities it makes sense to think subject to them.

6.1. The WKR Problem: Failed Solutions and an Alternative Approach

Rabinowicz and Ronnow-Rasmussen (2004, 2006) have convincingly argued that several attempts on behalf the FA-analyst to solve the WKR problem fail. To give the reader a sense of the difficulty of the WKR problem I will review the shortcomings of what I take to be three of the most natural attempts to solve it. I will then review a final, as yet problematic attempt to solve the WKR problem that I think is highly suggestive of the solution I will be proposing.

A first natural attempt to solve the WKR problem draws on Derek Parfit's (2001) distinction between what he calls "object given" and "state given" reasons for attitudes. Attitudes like desires and emotions have objects, or things they are about or directed towards. For instance, the object of a desire that a puppy is happy is a state of affairs in which he is happy, the object of one's guilt about lying is one's action of lying, and the object of one's anger at another person for lying is that person. Parfit calls a reason for an attitude "object given" just in case it is constituted by a fact about the attitude's object, while he calls a reason for an attitude "state given" just in case it is constituted by a fact, not about the attitude's object, but the state of one's having the attitude itself (Parfit 2001, 21-22). For instance, one's reason to desire a state of affairs constituted by the fact that it

involves happy puppies is object given. On the other hand, one's reason to desire a state in which one has an even number of hairs constituted by the fact that a demon will harm one's loved one's if one fails to have this desire is state given. Since the former but not the latter kinds of reasons seem to be those that constitute the goodness of a state of affairs, it is natural to try to solve the WKR problem by identifying the reasons of which FA-analyses of ethical concepts speak as object rather than state given reasons.

Rabinowicz and Ronnow-Rasmussen (2004, 406-7) rightly point out that the instantiations of intuitively "Cambridge properties," like *being the object of a desire which is such that if one has it one's loved ones will be spared* seem to give us an object-given reason corresponding to each state-given reason and vice versa. To make this criterion work we would seem to need a way to restrict the relevant reasons to instantiations of non-Cambridge properties. But a deeper problem that Rabinowicz and Ronnow-Rasmussen raise for this attempt to solve the WKR problem is that we can have reasons to have attitudes, which do not seem to contribute to the instantiation of a corresponding ethical concept, but which are constituted by instantiations of intuitively *non*-Cambridge properties of the object of the attitude. Suppose that if a Greek deity detects that you are angry at him for engaging in homosexual intercourse, he will feel that he is worthless. As a result he will work very hard and effectively to end world poverty in order to vindicate himself. Intuitively, the kind of reason one has to be angry with the deity for engaging in homosexual intercourse constituted by the fact that he has a disposition to end world poverty if you are does not contribute to the blameworthiness of his sexual act. But this reason certainly seems to be constituted by a fact about the deity who is the object of one's anger, or his instantiating a property that is at least as non-Cambridge as the property one's anger has of being such as to trigger his disposition to end world poverty should he detect it.

There is, however, still a way in which reasons like the above reason to feel angry at the deity relate to one's having the attitudes they are reasons to have, which might seem to distinguish them from fittingness reasons. While these reasons might be constituted by facts about the attitude's object, they still seem to "bring in" or be in part about the attitude they are reasons to have. A natural revision of the attempt to explain the difference between fittingness and non-fittingness reasons in terms of object as

opposed to state given reasons is thus to attempt to explain it in terms of reasons that do not, as opposed to reasons that do, mention the attitude they are reasons to have. I think that this is essentially the solution proposed by Jonas Olson (2004). Rabinowicz and Ronnow-Rasmussen (2006) point out, however, that while all non-fittingness reasons may well mention the attitudes they are reasons to have, some fittingness reasons do so as well. They observe, for instance, that the fact that someone is indifferent to being admired may well be the kind of reason to admire her that contributes to her being admirable. Indeed, the fact that someone is indifferent to the very token of admiration that I am thinking of having might well be this kind of reason that contributes to her admirability.

One final attempt to solve the WKR problem along these lines is due to Rabinowicz and Ronnow-Rasmussen (2004) themselves. As should be clear from our above case of feeling angry at a deity for engaging in homosexual sex, our attitudes are not simply directed towards things like persons, but towards such things for or on account of certain features of them.[157] One thing that might seem to differentiate one's reason to feel angry at the deity when this will end world poverty from reasons that contribute to his blameworthiness is that the former seem to involve his having properties that are other than those it is a reason to feel angry at him for having. Rabinowicz and Ronnow-Rasmussen consider the general proposal that fittingness reasons to have an attitude towards an object for having a set of properties just are the object's instantiation of any of these properties, and reasons constituted by the object's instantiating other properties are non-fittingness reasons. Unfortunately, Rabinowicz and Ronnow-Rasmussen also show how this proposal is open to counterexample. Let us alter the case of the Greek deity and suppose that he will not be moved to end world poverty unless you feel angry at him, not for having engaged in homosexual acts, but for being such that he will respond to anger by ending world poverty. The fact that he will end world poverty if one is angry at him on this score would thus seem to count in favor of being angry at him *for being such that*

[157] Perhaps our desires that states of affairs obtain are not simply directed at states of affairs, but rather at certain features of them, though this might seem more doubtful. Our desires for states of affairs certainly arise due to certain of their features, and we certainly take certain of their features to provide reasons for desiring them, but it would be another thing for certain features of the states, rather than simply the states, to be part of what they are desires for or about. For this reason it is doubtful whether this kind of solution could aspire to distinguish fittingness form non-fittingness reasons to desire that states of affairs obtain.

he will end world poverty if one is angry at him. Although this consideration thus satisfies Rabinowicz and Ronnow-Rasmussen's proposed criterion for a fittingness reason, it does not seem to contribute to such anger's being fitting, or to the blameworthiness of the deity's disposition.

While these problems with such natural attempts to solve the WKR problem might seem disheartening for FA-analysts, some might think we were too hasty to allow that there is a genuine problem here, at least as Rabinowicz and Ronnow-Rasmussen understand it. Both Gibbard (1990, 36) and Parfit (2001, 27) insist that what I have been calling 'non-fittingness reasons' or 'reasons for attitudes that do not contribute to the instantiation of ethical concepts' are not in fact reasons for such attitudes at all. Gibbard and Parfit insist that these are merely reasons to want to have or try to get oneself to have the relevant attitudes. Thus, the fact that a demon will harm one's loved ones unless one desires that one has an even number of hairs is not a reason to desire that one has an even number of hairs, but simply a reason to want to or try to get oneself to have such a desire. Similarly, the fact that a deity will end world poverty if one is angry at him is not a reason to be angry with him, but rather a reason to want to or try to get oneself to have such anger.

Rabinowicz and Ronnow-Rasmussen (2004, 412) agree with Gibbard and Parfit that the considerations in question are reasons to want to and try to get oneself to have these desires and emotions, but they do not share the intuition that it is inappropriate to describe them as reasons for the desires and emotions as well. Much more importantly, Rabinowicz and Ronnow-Rasmussen note that whatever we decide to call the reasons of which FA-analyses do not intend to speak, the FA-analyst needs to explain what makes a consideration a member of this category in terms that do not reference the ethical concepts she is trying to analyze. Even if Gibbard and Parfit are correct, there is still a threat to the FA-analyst. This is that what makes a consideration a reason to, say, desire a state of affairs instead of a mere reason to try to get oneself to desire it might *just be* that it contributes to the state's goodness (Rabinowicz and Ronnow-Rasmussen 2004, 413-14). The Gibbardian or Parfittian FA-analyst is still in need of a way to distinguish

reasons for attitudes from reasons to get oneself to have attitudes that does not invoke the concepts she is trying to analyze.

Even if one does not initially share Gibbard and Parfit's linguistic intuitions, I think that it is highly instructive to ask why it is that they seem attracted to a description of non-fittingness reasons, not as reasons for the attitudes in question, but merely reasons to want to have or get oneself to have them. I think that an important part of the answer is that, as we saw in chapter 1, judging oneself to have reason to have a given kind of response can (and typically does) directly cause one to have it. Judging that one has reason to desire a state of affairs because of such things as its involving happy puppies seems to be capable of causing one to desire it directly, or without one's having to do anything to get oneself to desire it.

But simply judging that a demon will harm one's loved ones if one doesn't want to have an even number of hairs does not seem capable of directly causing such a desire or feeling of anger. To respond to this second kind of reason one will have to do things like take pills, classically condition oneself, or rationalize oneself into thinking the attitudes fitting. What one's judgments about such reasons directly cause without behavioral intervention are only desires that one have the desire and behaviors undertaken in order to get oneself to have it. Thus, if one is attracted to the idea that judgments about reasons for a kind of response must be capable of causing one to have it directly or without one's having to do anything to bring it about, one might want to describe these other kinds of reasons as mere reasons to want or try to get oneself to have such desires and feelings of anger.

It might be attractive, then, to try to use these observations about the causal properties of judgments about various kinds of reasons to construct a solution to the WKR problem. We can in fact adopt this sort of solution whether or not we agree with Gibbard and Parfit's linguistic intuitions and whether or not we insist that all kinds of reasons for a response are capable of directly causing one to have it. We can simply try saying that to judge that an attitude is fitting is to be in a state capable of directly causing one to have it, while merely judging that one has non-fittingness reasons to (get oneself to) have it is not capable of causing one to have it without one's doing something to bring

it about that one does. Whether or not we agree with Rabinowicz and Ronnow Rasmussen that both judgments are about reasons for the attitude in question, we have a kind of criterion for distinguishing fittingness from non-fittingness reasons. This criterion certainly seems to categorize the above examples correctly, and it does not draw on the ethical concepts the FA-analyst is trying to explain. As we have seen, the criterion involved here would seem to correctly distinguish fittingness from non-fittingness reasons for attitudes outside the ethical realm. It certainly seems, and has been noted (see e.g. Kavka 1983, 36), that a feature that distinguishes evidential or epistemic reasons for belief from mere pragmatic reasons for belief is that judgments about the former can, while judgments about the latter cannot, directly cause one to have beliefs without behavior or activity undertaken in order to get oneself to have them.

6.2. A Particular Process of Direct Influence

As attractive as it might seem to explain the difference between judging an attitude fitting and judging oneself to have non-fittingness reasons to (get oneself to) have it in terms of the former having and the latter lacking direct causal powers, the approach faces important problems. As we mentioned, there are several ways in which judgments about *non*-fittingness reasons for attitudes can play a role in causing one to have them without one's having to do anything to bring it about that one does. To start with, consider the following kind of case:

> Suppose that Scott is convinced that he has been culpably wronged by Bill, who has taken his lunch money. Although he realizes that he would be justified in being angry at Bill, he doesn't at first feel angry. He rather just feels hurt and powerless. But then the thought occurs to him: "What kind of a little wimp am I being? You're not angry at him? Have a little pride!" This consideration immediately causes him to feel the anger that he already knew would be justified.

Intuitively, the fact that it would be cowardly to fail to feel anger at someone (or whatever fact or facts make it the case that this is cowardly) is not a consideration that contributes to the fittingness of the anger, or to the blameworthiness of the actor to whom it is felt.[158] But here we see that one's judging that one has this kind of reason seems capable of causing one to feel angry without one's having to do anything in order to get oneself to feel anger. So this kind of direct causal influence does not seem to be what separates judgments about fittingness from judgments about non-fittingness reasons for attitudes.

I think that this case is an instance of a more general phenomenon, namely that judgments that feeling *F* would be estimable (or disestimable not to feel) can directly cause one to feel *F* in a way that is similar in important respects that in which judgments that *F* is fitting can do this. As Velleman (2002) discusses, there seems to be an intimate connection between esteeming an ideal and coming to have its attitudes without any goal-directed processes of getting oneself to have its attitudes:

> The desire to mold oneself in the image of a generous person will meet with better success if it motivates one first to imagine being a generous person and then to enact this self-image, making believe that one is generous and using as props whatever motives one has the can be cast in the role of generosity. (Such props might be drawn from that fund of tenderness that Freud calls the Libido.) Emulating generosity in this fashion, one comes closer to being and to feeling generous, and one has a better chance of really becoming generous, by gradually working one's way into the role.
>
> …Now consider an attitude like respect or admiration for the ideal. Precisely because these attitudes are not goal-oriented motives, they tend to favor wishful thinking over purposeful activity. Admiring someone isn't a motive for bringing about anything in particular, and so it doesn't call for an instrumental calculation of the steps required to bring anything about. Wishfully picturing oneself in the image of an ideal is not a distraction from the business of admiring him: it is the business of admiring him. Emulation therefore flows directly out of admiration (Velleman 2002, 100-101).

In particular, judging that feeling *F* is estimable directly causes one to esteem it. But, as Velleman discusses, esteeming feeling *F* involves emulation, or acting as if and wishfully imagining that one felt *F* too, which can directly cause one to feel *F* without any goal-directed processes of getting oneself to feel it. Similar remarks go for disesteeming

[158] The fact that failure to feel such anger at someone might be *evidence* that what he did was culpable, since it may not be cowardly to fail to feel anger at anything but genuinely blameworthy actions. But this is distinct from the cowardliness of not feeling anger towards someone metaphysically *making it the case* that his action was blameworthy.

failing *F* and coming to feel it. The processes at work here seem similar to those documented by Philip Zimbardo (2007) in his 1971 Stanford Prison Experiment, where ordinary male college students (not to mention Zimbardo himself) came to genuinely possess the attitudes characteristic of guards and prisoners just by playing at the roles. This mechanism makes it the case that, as Strawson (1968, 90) noted, "Rehearsals insensibly modulate towards true performances."

Now, in the above kind of case the influence of one's judgment that failing to feel anger is disestimable on one's coming to have anger depended upon its first directly causing one to disesteem not feeling angry, and this disesteem's causing one to feel angry without one's having to do anything to get oneself to have the anger. So to take care of this kind of case we could simply amend the causal criterion on when a judgment about a reason is a judgment that it contributes to the fittingness of an attitude. The revised criterion would be that judgments that an attitude is fitting, unlike judgments that it is supported by non-fittingness reasons, exert direct causal pressure on one's coming to have them without one's having to do anything to get oneself to have them, and without first causing one to have other valenced attitudes like esteem for having or disesteem for not having the attitude.

But unfortunately for this quick fix there are other mechanisms by which what look like judgments about non-fittingness reasons to have attitudes can cause one to have them without one's having to do anything to get oneself to do so. It certainly seems that with external apparatus or neurosurgery we could create a case in which an agent's making judgments about any reasons to (get himself to) have attitudes we please would directly cause his coming to have them, without these *ipso facto* becoming judgments about reasons that contribute to the fittingness of the attitude in question. But we probably need not even move to such distant possibilities – actual world psychic mechanisms such as those involved in classical conditioning and wishful thinking would probably suffice. The development of sufficient mental associations between one's having an attitude with the obtaining of good consequences or the non-obtaining of bad consequences might well be enough to cause one to have it in response to circumstances that hold out prospects of such rewards.

Now, to be sure there still seem to be differences between the way in which judging an attitude fitting causes one to have it and the operation of the other kinds of actual world mechanisms we discussed above. Judging an attitude estimable or disestimable not to have will usually take a period of time over which (if Velleman is right) certain mental simulations take place in order for it to cause one to have the attitude. Similarly, forming sufficient emotional associations between, say, having a certain attitude and avoiding demonic threats for the former to be triggered by one's judgments about the presence of the latter would usually require a sufficient number of conditioning trials over a period of time. But judging an attitude fitting seems to be capable of causing one to have it without any such delay of time, process of mental simulation, or history of conditioning trials. [159]

Another difference concerns what happens when the causal influence of the mechanism in question fails to be decisive. When we judge an attitude fitting but fail to have that attitude, we experience a species of what psychologists refer to as "cognitive dissonance."[160] This cognitive dissonance is distinctive, however, in that it does not feel as though one is of more than one mind about an issue or as though one is being pulled in opposite directions. If anything, it feels rather like one is weak, inadequate to the demands of reason, or somehow deficient. Since this kind of cognitive dissonance arises

[159] This is so even if the conditioning mechanism by which we came to associate attitudes with things like demonic threats were to work like so-called "bait shyness," in which we become averse to a kind of food after only one subsequent bout of illness (see e.g. Zimbardo and Weber 1997, 210-211). Such a mechanism would still require one conditioning trial to operate, which marks a contrast between its causal influence and that of fittingness judgments.

[160] A complication here concerns whether we judge the attitude to be merely justified (which is the attitude we often take to anger as a response to blameworthy acts) as opposed to rationally mandatory, or irrational not to have (which as we saw we might often think is true of preferences for better states of affairs and feelings of obligation not to do wrongful things should one not already be sufficiently motivated to omit them). When one merely judges an attitude fitting in the sense of justified, one need not experience this kind of cognitive dissonance should one fail to have it. The kind of cognitive dissonance in question is present when one fails to have attitudes one judges to be rationally mandatory. There may, however, be circumstances in which the absence of an attitude in the presence of a judgment that it is justified but not mandatory can engender recalcitrance dissonance. This would likely be so if the (inconclusive) reasons in support of the attitude are those upon which one forms intentions to act by means of the "sucking it up" pathway that we saw in Chapter 3. Thus one might form an intention to go and play Go on the strength of the reasons that make it rationally optional to desire to play Go, one is apt to regard one's lack of inclination to play Go as recalcitrant. In the same way, if one forms an intention to save Suzy on the strength of one's reasons to feel obligated to save her (where such reasons to save her and Bugsy were equally matched), but fails to feel so obligated (one might, for instance actually hate both Bugsy and Suzy, but know that one still has to do one's best to save them), one might well regard one's lack of feeling obligated as recalcitrant.

in response to profiles of attitudes that are recalcitrant to or out of line with one's views about what attitudes to have, I shall call this kind of cognitive dissonance *recalcitrance dissonance*. Judging an attitude fitting seems to be distinct, then, in that when its causal influence fails to be decisive and determine that one has it, one feels recalcitrance dissonance. One does not experience such recalcitrance dissonance when mechanisms like disesteem for not having an attitude or classical conditioning fail to be sufficient to cause one to have it. When one judges that one has reasons with respect to an attitude like that one is a coward for failing to have it or that a demon will do bad things if one doesn't have it, and such mechanisms are insufficient to cause one to have the attitude, one may have various attitudes in response, ranging from disesteem towards oneself to fear about what will happen. But one will not experience recalcitrance dissonance.

I suspect that we cannot directly use any of these differences between the causal influence of fittingness judgments on our attitudes and the causal influences of such actual world mechanisms as esteem and classical conditioning in yet another quick-fix of the attempt to explain the difference between judgments about fittingness and non-fittingness reasons in terms of their direct as opposed to behavior mediated causal powers. I suspect that with enough neurosurgery, external machinery, and distant-from-actual psychology we could construct cases in which judgments about non-fittingness reasons have the above features that only judgments about fittingness reasons typically actually do. What I think we can do is use these features that distinguish fittingness judgments from psychic mechanisms like disesteem for failing to have an attitude and classical conditioning to hone in on an actual psychic mechanism that is involved in judging an attitude fitting. I think that we can then use this mechanism to explain the difference between judging an attitude fitting and judging that one has non-fittingness reasons to (get oneself to) have it.

6.3. Norm Acceptance

As I mentioned, I think that fittingness judgments' mechanism of attitude causation is best understood in terms of a mind state we may call *norm acceptance*. We can naturally

talk about accepting norms for attitudes like beliefs and credences. We can say, instance, that we accept *modus ponens* as a norm of deductive inference, that we accept norms that prescribe *ceteris paribus* believing the simplest explanation of a phenomenon, and that we accept norms that require us to conform our degrees of belief to the axioms of the probability calculus. When we speak in this way about accepting norms for belief, we seem to have in mind a state that is neither a belief about how our beliefs are nor a desire about how we would like them to be. Rather, just as beliefs and desires are states with the functional roles of combining to directly produce behavior, these states of norm acceptance have the functional role of directly revising attitudes like beliefs and desires.

There are at least two senses in which we might speak of someone accepting a norm for an attitude. On the first, to accept a norm just is to have a distinctive kind of tendency to conform to what the norm *actually* prescribes, rather than what you merely take it to prescribe. In this sense of accepting a norm, if you accept a norm of the form *Believe P if Q!* or *Feel F if C!*, then you will have a direct propensity to believe *P* or feel *F* if you *Q* to obtain or you take yourself to be in *C*. There is no psychological fact of the matter about what the norms you accept actually prescribe beyond the patterns of attitudes they actually tend to cause you to have. You can of course be mistaken about what norms you accept in the same ways that you can be mistaken about what you believe of desire. But as with beliefs and desires, it is what you actually shallowly accept, rather than what you represent yourself as shallowly accepting, that plays the psychological causal role of regulating your responses. In the shallow sense of norm acceptance, there really is no room to maintain that although someone takes himself to be in circumstances of kind *C* and has no propensity to feel *F*, that he "really, deep down accepts" a norm that prescribes feeling *F* in circumstances *C*. For this reason I shall call this first kind of norm acceptance *norm acceptance in the shallow sense*.

But to accept a norm in the shallow sense still requires that one is in a state that plays distinctive functional roles. One role of this state is of course to directly influence attitudes like beliefs, desires, and emotions much in the way the latter states play the role of directly influencing intention and behavior. Another distinctive feature of such states is that the failure of their causal influence to be decisive gives rise to what I above called recalcitrance dissonance. A final and important feature of states of norm acceptance is

their ability to combine with each other to constitute a subject's accepting a system of norms.[161] As our examples of epistemic norms should make clear, we do not accept only one norm for a kind of attitude; our norms for attitudes form a system that works together to prescribe a response. A state of shallowly accepting a norm can thus exercise its influence by strengthening or inhibiting the influence other norms, causing new states of norm acceptance to come into existence, and causing other states of norm acceptance to cease to exist.

It might seem, however, that we can speak of accepting a norm in such a way that one can fail to correspond to the prescriptions of norms that one genuinely accepts, even without recalcitrance dissonance. First, there are cases in which a norm we accept prescribes a response but we fail to recognize that it does so. We might, for instance, believe both:

Q: Quantum mechanics is true, and
M: If Quantum mechanics is true, then there are many worlds.

Although we accept *modus ponens*, we may fail to infer that there are many worlds from *Q* and *M* by simply overlooking *ponens*' applicability to our situation – say because we fail to "put *Q* and *M* together."

Second, there are cases in which we are caused to have a response in much the same way as when we recognize that our norms prescribe it, but where the response is not in fact prescribed by norms we accept. Consider cases of fallacious deductive inference. We might, for instance, believe that:

C: If society should be Communist, then we should redistribute income,
F: Society shouldn't be Communist,

and infer from *C* and *F* that:

[161] And perhaps even more complex structures like her ruling out sets of systems of norms. On the notion of ruling out systems of norms, see (Gibbard 1990, Chapter 5).

R: We shouldn't redistribute income.

Although we can make inferences like this, it certainly does not seem that we actually accept norms that prescribe denying the antecedent. In the case of such inferences, it seems that a mere appearance that *C* and *F* commit us to believing *R* causes us to believe *R*. But the direct influence of such erroneous appearances of commitment seems identical to that of appearances that correspond to the prescriptions of norms we do accept, like that to the effect that *Q* and *M* commit us to belief in many worlds. Indeed, if it looks to us like *C* and *F* commit us to *R* and we fail to believe *R*, we will tend to have the same kind of recalcitrance dissonance that we have when we fail to believe in many worlds despite its looking like *Q* and *M* commit us to such belief.

The fact that we can fail to conform to the norms we accept in these ways just described might suggest that our attitudes are governed by *representations* of what is prescribed by norms that we accept in some sense that is deeper than just what we tend to conform to absent recalcitrance dissonance. These representations can be mistaken, and when they are we have a tendency to conform to what we represent our norms as prescribing rather than what they actually prescribe. If this is right, then we must have a way of representing the norms we accept without being able to know all of their prescriptions. At the same time, the representations of our norms that play a role in inference surely do not have a mode of presentation like THE NORMS I ACCEPT, WHATEVER THEY ARE. Perhaps the mental states that represent what our norms prescribe do so in virtue of bearing a certain nomic relation to the mental states that constitute our acceptance of these norms. Candidates for this nomic relation might resemble the kind of information carrying under ideal conditions discussed by Stampe (1977) and Dretske (1981), or the kind of asymmetric causal dependencies discussed by Fodor (1987, 1990).

To speak of norm acceptance in this sense is to speak of an underlying state, representations of which have a propensity to alter our attitudes in accord with what they represent our norms as prescribing. For this reason I shall call this kind of norm acceptance *norm acceptance in the deep sense*. States of deep norm acceptance are still distinctive in that they form a system that directly influences our attitudes and give rise to

recalcitrance dissonance when the influence of this system fails to be decisive. But when we are talking about deeply accepted norms, the causal influence of the system of norms we accept our attitudes of belief, desire, emotion, and other states of acceptance norms is mediated by representations of it prescribes.

6.4. Norm Expressivism and its Discontents

The notion of the acceptance of norms for attitudes in both the shallow and deep senses give rise to some natural ways of explaining the direct influence that fittingness judgments on our attitudes. First, we can observe that to accept in the shallow sense a norm or a system of norms that prescribes an attitude in a particular circumstance looks an awful lot like judging it fitting to have that attitude in that circumstance. As such, we might just try analyzing the judgment that it is fitting for someone to have a given attitude as a state of accepting a system of norms that prescribes having it in her circumstances, giving us a version of what we might call the:

> **Norm Expressivist Theory of Fittingness Judgments:**
> To judge that it is fitting for some agent, *A* (where *A* is often the judge herself), to have attitude *F* is to shallowly accept a system of norms that prescribes having *F* in *A*'s circumstances.[162]

By giving an independent account of what it is to judge an attitude fitting in terms of the attitude of norm acceptance, Norm Expressivism gives us a way to solve the WKR problem. To judge it fitting for one to have an attitude is to accept norms that prescribe

[162] This sort of Norm Expressivist account is very close to that which we could derive from Gibbard's (1990, 86-92) "second approximation" of Norm Expressivism. A version corresponding to Gibbard's final view would need to make reference to something like ruling out sets of ordered pairs of complete systems of norms and possible worlds (see Gibbard 1990, 94-99). As Gibbard points out, this final version of Norm Expressivism can better handle several problems the second cannot, including that of giving a semantics of normative language in embedded contexts that can solve the Frege-Geach problem. It is for the sake of simplicity and ready comparison with other views I discuss that I explicitly discuss the second approximation. My second approximation talk could be replaced with third approximation formulations without any substantive effect on what I have to say.

it. To judge the attitude to be supported by non-fittingness reasons, like 'the demon will harm my loved one's if I don't have the attitude' or 'I'll be a coward if I don't have the attitude' is to be in a different state of mind. Most plausibly this different state is one of accepting norms that prescribe being motivated to get oneself to have the attitude or disesteeming not having the attitude. This would explain why these other judgments exert the direct causal influence characteristic of norm acceptance on attitudes like motivation to get oneself to have an attitude and disesteem for not having the attitude, but not the attitude itself. If we like we can still follow Rabinowicz and Ronnow-Rasmussen and call these judgments that one has reasons for the attitude in question. But we will have found a way to distinguish them from fittingness judgments, in a way that explains which attitudes they are incapable of directly affecting, and in a way that does not invoke the ethical concepts the FA-analyst is trying to analyze.

Now for reasons we saw in chapter 3, judgments about what valenced attitudes are fitting are judgments about what ends to pursue, which, when we add to them judgments about the fittingness of the relative strengths of our intrinsic motives, determine a set of views about the entirety of our objective reasons for action. On the Norm Expressivist account, this translates into the idea that to make a judgment about objective reasons for action is to make a judgment about what will actually best satisfy the intrinsic motivations prescribed by one's system of norms.

But Norm Expressivism can be more than just a theory of fitting attitudes and objective reasons for action; it can be cast as a theory of all normative concepts. Indeed, the considerations that make it attractive as a theory of fittingness assessments and objective reasons for action make it equally attractive as a theory of other normative notions. Thus, to judge that one has epistemic reason to believe something is to accept a system of norms that prescribes belief in it, and to judge that a particular claim about decision theory or instrumental rationality is true (say that Causal Decision Theory is) is to accept a system of norms that prescribe intentions and actions given or in light of the prescriptions of one's norms for beliefs and valenced attitudes or intrinsic motivations. We can summarize this more general view by characterizing Norm Expressivism in the following way:

Norm Expressivism:

To judge that agent *A* should ϕ (where ϕ can be any response – believing, feeling, intending, or doing something) is to shallowly accept a system of norms that prescribes ϕ-ing in *A*'s circumstances.[163]

According to Norm Expressivism, our normative judgments are not descriptive or representational states. This may well seem to be an advantage in that it seems to capture the way in which fittingness assessments and other normative judgments play a role in guiding our attitudes and actions that is more intimate than any other representational states we might think of. It also seems to make sense of the apparent metaphysical and analytic independence of the normative from the descriptive. As we saw in chapter 1, it does not seem that we need analytically non-reducible normative facts to explain our beliefs about them. But this does not seem to show us that normative concepts are empty. The apparent explanatory impotence of normative facts does not seem to mean that no considerations really count in favor of beliefs, valenced attitudes, or actions, or that every belief, action, desire, and emotion is as poorly justified or supported by reasons as every other. At the same time, it does not seem that normative judgments are analytically reducible to anything that we do need for explanatory purposes. It seems coherent, even if wildly implausible, for people to think that just about any descriptively understood consideration is a reason to do, think, or feel something. Attempts to analytically reduce normative judgments to descriptive or representational states thus seem to clash with our intuitions of coherence.

According to Norm Expressivism, this sort of metaphysical and semantic independence of the normative is exactly what we would expect. If normative judgments are nothing but non-descriptive mental states of norm acceptance, there is no obvious reason why we would treat the explanatory irrelevance of the truth of these states as a reason against being in them. In the same way, it would be obvious why we have a boundless array of coherent judgments about the normative import of a given descriptive

[163] Here the thought that one should believe or feel something is restricted to the thought that the attitude is warranted: fitting or supported by epistemic reasons. To get to the thought that one "should believe / feel" in other senses we apply the analysis seen above, interpreting it as the acceptance of a system of norms that prescribes thing like esteeming having the attitude, wanting to have the attitude, or getting oneself to have the attitude.

feature. If normative judgments are states of accepting norms for what to believe and feel in a given situation, there will be nothing incoherent about states that have any given prescriptions for what to believe or feel in the presence of any given descriptive features. Moreover, it seems very difficult to explain the metaphysical and epistemic independence of the normative in any other way. If our normative judgments were true or false in virtue of corresponding or failing to correspond to normative facts of the kind that can explain or fail to explain why we hold the normative judgments we do, it looks very hard to see how epistemic reasons to believe in them could not be tied to explanatory considerations of parsimony. This is what holds for facts about Zeus, atoms of phlogiston, and in other areas of philosophy it is the serious standard by which proposals about abstract entities are assessed and worried about.[164] Why should the standard for descriptive normative facts be any different? Chiefly, we feel, because normative facts just don't need to play certain explanatory roles in order for us to be justified in thinking that they obtain. But the fact that this is so, yet explanatory roles do seem relevant for every other kind of descriptive fact, is exactly what expressivism stands to best explain.

As attractive as Norm Expressivism might thus be, it often tends to arouse a sense of dissatisfaction. I believe that the sense of dissatisfaction is out of line with the merits of the view, but that it points to a major problem for it. People try to express their dissatisfaction with views like Norm Expressivism by saying that these views cannot make sense of how there are normative truths, or normative facts, or how these are objective and real and that we can get them right or wrong. But with the advent of the quasi-realist program, expressivists can embrace minimalism about truth and facthood, and they can explain talk of objectivity as normative talk expressing norms that have prescriptions for people's situations that are insensitive to what their beliefs and desires are like. But somehow this does not quite seem to get to the heart of the issue. The heart of the issue, I think, is that Norm Expressivism has a difficult time making sense of what we are doing when we engage in deliberation about what to do, think, and feel.

It is far from straightforward what kind of account of basic first-personal normative inquiry can be given by Norm Expressivism, which identifies judging that one

[164] See for instance Putnam (1972), Benacerraf (1973), and Quine (1980).

should do, think, or feel something with simply shallowly accepting a system of norms that prescribes so doing, thinking, or feeling it. Norm Expressivism seems to be motivated largely by an attempt to make sense of interpersonal normative discussion, and actually seems to leave little room for understanding what goes on when we try to figure out for ourselves what to do, think, and feel, as opposed to simply showing up to normative discussions with a more or less complete system of norms, ready to bend others to it or get it bent to the norms of others.[165] In light of Norm Expressivism's focus, it is in fact not too surprising that the very few things Gibbard (1990) says about first-personal normative inquiry in his systematic development of Norm Expressivism seem to be attempts to explain it as a kind of rehearsal for interpersonal normative discussion:

> To prepare oneself to meet demands for consistency [i.e. in normative discussion] may require a strong imaginative life. A person will engage in imaginative rehearsal for actual normative discussion; he practices by holding himself to consistency. The pressure for consistency need not be so strong as it is in good philosophical discussion, but it will be there, and it may be significant (Gibbard 1990, 74-75).

> In trying to decide what is rational, we are engaging our normative capacities to try to decide what norms to accept. We do this in normative discussion, actual *and imaginative*, as we *take up positions*, subject ourselves to demands for consistency, and undergo mutual influence [emphasis added] (Gibbard 1990, 81).

The picture Gibbard seems to be suggesting is that when we engage in basic normative inquiry into what to do, think, or feel, we imaginatively take up the positions of agents who accept different systems of norms, that these systems of norms alter on contact in the way he suggests they do in flesh and blood normative discussions, and that we emerge with the system of norms that ends up being shared by the parties to the normative discussion in our heads.

I fear, however, that it is somewhat unclear how the details of this account are supposed to go, and that the more one tries to fill them in, the less it looks like a plausible account of normative inquiry. How exactly are we to take up these systems of norms and be influenced by a simulated discussion between their proponents? It would seem

[165] It is interesting to note that W.D. Falk (1963, 1986) claimed that in his day expressivist metanormative theories seemed primarily developed in order to account for the dynamics of normative discussion, which caused them to face problems in accounting for first-personal normative thought generally, and in particular deliberation about what to do, think, and feel. In this respect little may have changed.

fantastic to suppose that for a time our psychie actually fractures into parts that fully accept different systems of norms. The mere fact that one engages in normative inquiry does not seem to entail that one is subject to conflicting pressures on one's attitudes. Normative inquiry usually seems to involve suspension of judgment about what to do, think, or feel rather than one's holding conflicting views – indeed it often does not seem to require that one have much of a sense of what the competing options might be or what they might have going for them.

Perhaps, though, the account does not demand that our psychie split into bits that each accept different systems of norms. Perhaps it is enough that we simply imagine having a normative discussion with characters we think of as accepting a certain systems of norms. But it seems difficult to see how the kind of brute influence on the system of norms one accepts that Gibbard talks about would be exerted just by imagining talking to people with any systems of norms we choose to imagine up. Perhaps for the exercise to work one needs to be imagining not merely any stipulated person but someone who in some way resembles agents with whom one has had or is likely to have normative discussions. I think that this kind of imaginative rehearsal for normative discussion really does take place, for instance in planning for political debates or presenting material to students. But this looks a lot more like deliberating about how to be most rhetorically or pedagogically effective in convincing someone of conclusions one already accepts (or in the political debate case, just coming out looking good and making the opponent look stupid).

Of course, this kind of imaginative rehearsal for debate or presentation can sometimes *prompt* genuine deliberation about what to do, think, and feel - for instance if one comes to a point in the rhetorical or pedagogical planning about which one realizes one is not really sure oneself. But to the extent it does one seem to stop engaging in an imaginary normative discussion and to start doing something else, which is not essentially tied to rehearsal for discussion. Perhaps one could use images of the same characters with whom one was planning to discuss things as stooges who one imagines giving voice to some of one's doubts or counter-arguments one can think of, but then these imagined characters are just that – voices for considerations one is thinking up oneself, which one need not really imagine them voicing.

I do not think that it is an accident that Norm Expresivism has such a difficult time explaining what we are doing when we engage in basic normative inquiry into what to do, think, or feel. As we saw in chapter 1, there are features of this kind of normative inquiry that seem to cry out for a descriptivist treatment of normative thought. Often times, we will have an intuition or set of intuitions – say that we owe a great deal to those suffering right in front of us and that we owe very little to those who are far away. We see then that they conflict with our other intuitions, for instance that bare physical distance does not affect the strength of our moral obligations. In trying to decide what to do, we try to see which of these intuitions is the result of a distortion, or of being thrown up by something other than an accurate perception of the normative facts. This looks on the face of it exactly like the sort of thing we do when we try to see if our perceptions have been generated by explanatory intercourse with what they represent.

Now the Norm Expressivist can of course offer an alternative explanation of this talk about distortion, according to which it is in part normative rather than purely metaphysical in character. It is somewhat unclear what the expressivist might want to say about normative intuitions, but that is no serious problem here; perhaps the expressivist thinks that they are just relatively spontaneous tendencies to accept norms. The expressivist might say, then, that what we are trying to do in deciding whether an intuition is veridical is that we are trying to decide whether it is the result of a process that our epistemic norms tell us to trust, or to weigh in favor of accepting the norm that it is a tendency to accept. One problem for this picture, however, is that often (perhaps even most of the time) we do not resolve the dispute by learning new empirical information about how our intuitions were formed. We usually elicit more intuitions, with content that is both general and particular. Does the view that distance matters get us into more trouble with case intuitions? Could there be something other than distance *per se*, which is at least as intuitively irrelevant as we think distance is, the irrelevance of which we might be confusing with that of distance?

Of course, the expressivist might then offer this for an epistemic norm we all accept: weigh an intuition in favor of accepting the relevant norm iff the intuition is not explained away by something else. The problem here concerns what it is to explain an

intuition away. The standard way to describe this is that we show an intuition to be mistaken by showing that something other than the truth of its content explains why we have it. It might seem that such and so feature is normatively relevant, but we can see that the appearance is misleading on the grounds that the explanation of the intuition does not have to cite its truth. So for instance, one might attack the epistemic credentials of the intuition that we do not owe so much to distant strangers on the grounds that distance is thought to be correlated with feasibility, or personal relationships, or what have you, trying somehow to show that we can explain why we think distance matters in terms other than distance mattering. But this looks like our good old fashioned attempt to debunk the epistemic credentials of a representation by means of an explanation that does not cite its truth. In practicing this method we presuppose that not all of our intuitions can be explained in terms that do not cite their truth – unlike those we single out for debunking, we cannot explain those we continue to rely upon (for instance that we owe a great deal to nearby strangers) as the product of forces that are insensitive to its accuracy. Our reflective equilibrium methods thus really do seem to presuppose a kind of basic metaphysical sensitivity of our normative intuitions – *some* of our normative intuitions – to the normative truth, whatever exactly that normative truth consists in.

Indeed, as I mentioned, I think that the difficulties expressivism has explaining normative inquiry and the apparent metaphysical commitments of our reflective equilibrium methods can be subsumed under an even more general feature of the relationship between inquiry, knowledge, and explanation. In inquiring into whether P is true, we are trying to align our judgments about whether *P* is true with facts of the matter about whether it actually is true. But since inquiry cannot seek to achieve correspondence between our judgments about *P* and the facts about *P* by total accident, it must aim at establishing some kind of metaphysical relationship between *P*'s truth and our judgments about *P*. Another way to think about this would be to say that inquiry aims at knowledge, for whatever else knowledge is it is true belief that is not true by total accident. Knowledge thus requires some kind of metaphysical relationship between an instance of knowledge and the truth of the proposition of which it is knowledge. In particular, it seems that knowledge about *P* requires – and inquiry into whether *P* is true must seek to establish – a state of affairs where our judgment that *P* (if *P* is true) or not-*P*

(if *P* is false) is in part explained by the fact that *P* or the fact that not-*P*. But this is exactly the sort of metaphysical relationship that cannot obtain between normative judgments and normative facts if expressivism is true. That normative inquiry aims at establishing such a metaphysical relationship between normative judgments and normative facts would explain why Norm Expressivism seems unable to make sense of its epistemic point.

6.5. Deep Norm Acceptance and Basic Normative Inquiry

But if Norm Expressivism is dissatisfying on account of its problems explaining what we are doing when we engage in basic inquiry into what to do, think, and feel, what would a more satisfying account look like? Interestingly, a significant and rather diverse group of philosophers have been attracted to the idea that we use reflective equilibrium methods of normative inquiry to discover our own deepest commitments. The idea is that such methods seek to uncover the structure of an underlying "practice," "capacity," "sense," or "set of values" that generates our normative intuitions and judgments, but to which we lack conscious access.[166] Such methods can, however, cause us to extend and revise our normative views in significant and even radical ways. Many kinds of distorting influences, including confusion, wishful thinking, emotional biases, and faulty theorizing can cause our normative intuitions and judgments to fail to reflect our underlying commitments. In seeking a unification of our normative intuitions we attempt to determine which can be debunked as products of distortion and construct a theory of our commitments that best explains the intuitions they generate. In this way we achieve access to our genuine commitments in much the way empirical inquiry achieves access to the external world by constructing a best explanation of our perceptual experiences, which are causally sensitive but by no means infallible guides to it.

Now, since our normative judgments directly guide our attitudes in a way unlike representations of beliefs and desires, it seems that the reflective equilibrium methods

[166] See for instance (Goodman 1954), (Rawls 1971), (M.B.E. Smith 1977, 1979), (Fischer and Ravizza 1992), (Kamm 1993), (Unger 1996), and (McMahan 2000).

that generate these judgments cannot be attempts to discover what we believe or desire. But what if, rather than beliefs or desires, the underlying "sense" or "values" that we use reflective equilibrium methods to uncover are the norms that we deeply accept? As we have seen, we are governed by the norms we deeply accept by means of representations of what they prescribe, which representations can be erroneous. Yet quite independently of its veracity, a representation that a deeply accepted norm prescribes a response exerts direct causal influence on our coming to have the response, and should this influence fail to determine how we respond, we will tend to experience recalcitrance dissonance. Identifying reflective equilibrium methods of basic normative inquiry with inquiry into what our deeply accepted norms prescribe might thus seem to explain two of its central features. The first is that in practicing the method we are trying to align our views about a subject matter with the facts about that subject matter – just as we would be doing in any other kind of inquiry. We are, that is, trying to reach a state where the facts about our subject matter play a role in explaining our judgments about them. If there are facts of the matter about what the norms we deeply accept prescribe that are quite imperfectly tracked by appearances to us about what their content is, we will have on our hands a project of trying to get our views in line with what they actually prescribe. We can, that is, try to get our judgments about what is prescribed by the norms we really deep down accept to be caused by the facts about what we accept by going with the appearances that reflect them and by refusing to go with the appearances that do not.

The second thing about reflective equilibrium methods we could explain by identifying them with inquiries into what our deeply accepted norms prescribe is the way in which we are governed in accordance with our *views* about the facts we seek rather than the facts we seek themselves. As we saw, we behave in accordance with the beliefs and desires we actually have rather than the beliefs and desires we merely take ourselves to have. But we do not necessarily infer in accordance with the norms we deeply accept – both you and someone who makes fallacious inferences may deeply accept the same deductive norms, but your non-fallacious inferences are guided by accurate representations of what these norms prescribe while his fallacious inferences are governed by inaccurate representations of what they prescribe. Now it is surely the case that false normative views can be just as causally efficacious on our attitudes as true ones.

Thus if you used to think that you owed more to members of your own race but reflective equilibrium methods changed your mind, you might well have been guided by your racist views in the past just as much as you are guided by your egalitarian views now. In the past you would have had a propensity to feel more strongly obligated to help members of your own race, and you would have tended to experience recalcitrance dissonance if you didn't. Now you have a propensity to feel obligated to treat everyone equally, and you tend to experience recalcitrance dissonance when you do not feel that way.

So if we want to identify reflective equilibrium methods as inquiries into one's own deepest commitments, we had better be able to say that it's your views about what you are committed to rather than what you are actually committed to that determines how you respond and whether your experience recalcitrance. For it's not as though you were much of an egalitarian before you discovered your commitment to equality. You might have had impulses to treat people equally, surely – just as you might have impulses to discriminate against people now. If the discovery of commitment picture is correct, it appears that causal influence of normative governance is on the side of perceived commitment rather than actual commitment.[167] But this is exactly what we would have if reflective equilibrium methods are inquiries into what is prescribed by the norms that we deeply accept. For you might all along have deeply accepted a norm that prescribed feeling obligated to treat people equally. Your previous (inaccurate) representation that your deeply accepted norms prescribed no such thing would have been just as causally efficacious as your current (accurate) representation that your deeply accepted norms prescribe exactly this.

Given its potential to explain what goes on in basic normative inquiry, let us take a closer look at the picture of what it would be to deeply accept a norm. We may recall

[167] Moreover, while intuitions would be the primary source of information about our commitments in the relevant sense, a person's bare recalcitrant inclinations (which, to recall from Chapter 2, can be just as recalcitrant to her intuitions as they can be recalcitrant to her judgments) would indicate nothing. The bare fact that you feel guilt for knocking over a lamp is, in the absence of intuition, no evidence at all that you should feel that way. If you feel guilt that in every way appears unwarranted, and in no way warranted, you have nothing to suggest that you should be feeling guilty. A sneaking suspicion that you really should be afraid of something is some evidence that you should – the bare fact that you feel fear that seems nothing but phobic is not. A sneaking suspicion that it would be desirable to wash your hands is one thing – a transparently compulsive desire to do so is quite another.

from our discussion of shallow acceptance that accepted norms are in the first instance norms for attitudes like beliefs, emotions, desires, and intentions. According to the deep model of norm acceptance, there are three components involved in accepting a norm for an attitude: (1) the underlying state of norm acceptance, (2) representations of that state of norm acceptance, and (3) the attitudes of the kind that are governed by the norm. We can think of all of (1)-(3) as contentful mental states. The acceptance of the norm will have as its content (surprise, surprise) a norm that has prescriptions for the kind of attitude in question. We can think of norms as ways of assigning prescribed attitudes (including absences of attitudes and degrees or strengths of attitudes) to descriptions of the circumstances in which they are described.[168] Similarly, the representation of the accepted norm will have a proposition as its content, namely that the norm in question has such and so prescriptions for the attitude it governs. The content of this representation could in principle be as general as a claim about the entire content of the norm or as specific as its prescriptions for a single fully specified circumstance. Finally, the attitude governed by the norm will have whatever content we typically think of it as having, so for instance if it is a desire for a state of affairs it will have the desired state of affairs as its content, if it is a feeling of obligation it will have the thing one feels obligated to do or omit doing as its content, and so on.

By way of a purely hypothetical example, suppose that we deeply accepted a norm for when to feel obligated not to kill that was roughly of the form *Feel strongly obligated not to kill sentient beings who are psychologically connected to futures that are good for them!*[169] We might, however, have confused the property of being a sentient being connected to the relevant kind of future with a property with which it saliently co-

[168] Roughly, one might try to think of norms as mappings from circumstances to attitudes, or ordered pairs of circumstances and attitudes prescribed in them. But I suspect that this will be too coarse a notion for our purposes. For on the picture we are developing, representing a norm as one way or another involves having a view about what is normatively relevant, and we might want to distinguish between the view that one or another of two necessarily co-occurrent features is normatively relevant, or that one or the other of two ways of weighing different considerations accurately captures the strength of one's reasons, even if they must always prescribe the same responses.

[169] For discussions of the notion of psychological connection or continuity see for instance Parfit (1984) and McMahan (2002). But it should be noted that the relevant facets of psychological connection and continuity may involve quite a bit more of the less self-conscious aspects of our mental lives than these authors might seem to suggest. They might incorporate a great deal of connection and continuity constituted by the systems of episodic, procedural, and emotional memory responsible for the majority of our psychological functioning over time, as well as between the psychological dispositions constitutive of what is commonly referred to as 'personality'.

varies – namely that of being a human organism.[170] As a result, we may represent our deeply accepted norm as prescribing that we should feel strongly obligated not to kill all and only human organisms, including such human organisms as are not beings that are psychologically connected to futures that are good for them, like early embryos and people in excruciating and irremediable agony. If this happens, then we will tend to feel obligated not to kill all human organisms – embryos and people in irremediable agony included. The relationships between (1) the state of norm acceptance, (2) the representation of the norm's prescriptions, and (3) the feelings of obligation that result are depicted below in figure 8.

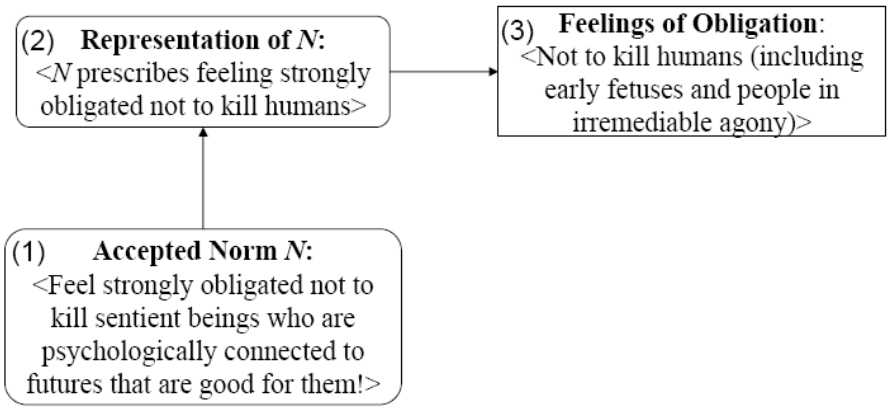

Figure 8: The Three Components Involved in the Deep Acceptance of a Norm for An Attitude (In this Particular Case, for Feelings of Obligation)

[170] Where I mean 'human organism' in the literal, biological sense – the only sense, by the way, in which anyone has any business talking. There is a certain tendency among some people to use the term 'human' as itself normatively loaded, as in 'What we need to ask about fetuses what is at what point they become human.' This terminology is as pernicious as it is absurd. It invites the grossest conflation between the biological criterion and whatever property it is that actually is relevant to how we ought to treat entities. When egalitarians argued that race and sex were irrelevant, they simply said that they weren't – they did not concoct an expanded sense of 'white' or 'man' and start saying that minorities and women are really white men. I am sure that space aliens would feel deeply insulted if we had a history of discriminating against them on the basis of species membership but then started calling them 'humans' in some expanded sense. I am well aware that some relatively egalitarian talk is couched in terms of 'human rights'. You might also remember there being times when *relatively* egalitarian talk was couched in terms of 'the rights of citizens' and 'the rights of men' (where they really did just mean property owning citizens and biological males). If you think that a category isn't really invested with significance, you're likely to do better not talking as if it is. (You will, for instance, be less likely to trick yourself into thinking or presupposing that the category is relevant after all. 'It'll be so much better for all Italians if we just pretend that *being Italian* is a normatively relevant property' – and how will it be for the non-Italians whose interests you are only "pretending" to be less important?).

Like psychological states more generally, I think that these states of norm acceptance, norm representation, and feeling obligated should be understood as certain functional states of the subject who has them.[171] Our states of norm acceptance and norm representation will presumably be states of our brains, but an alien being's states of norm acceptance might be states of the silicon-based stuff out of which he is composed, a ghost's states of norm acceptance might be states of his ectoplasm, and so on. What makes these states count as states of norm acceptance, norm representation, and feeling obligated is that they play a certain functional roles – that they interact with other states of the relevant kinds in ways that eventually produce behavior. In this sense the notion of being a state of norm acceptance or norm representation is relevantly like that of being a poison.[172] What makes a given chemical substance a poison is the fact that it has the ability to make some contextually salient being ill.[173] In the same way, what makes a neural or silicon state a state of accepting a norm for feeling obligated is its having the right propensity to cause states of norm representation which in their turn have the right propensities to cause states of feeling obligated. Neither state would count as the acceptance of a norm for feeling obligated or the representation of such a norm unless the first had the right disposition to give rise to the second and the second had the right dispositions to give rise to feelings of obligation.

Now while it is by no means uncontroversial, I shall assume that functional states like these can have intentional contents, and that what makes it the case that they have the contents they do is something about their particular functional roles. Start here with the state of norm representation. There are several proposals on offer about how a functional state might get to have a certain representational content. The one I find most plausible and germane to my purposes here is essentially some version of Fodor's (1987, 1990) theory of asymmetric causal dependence. I shall thus describe things in terms of an

[171] While my sympathies are with analytic functionalism across the board, this is not intended to beg any hard problems in the philosophy of mind. You can, if you like, read my talk of 'psychological states' as referring to those Chalmers (1996) distinguishes from phenomenal states or states of subjective experience. If so, then any phenomenal elements of the valenced attitudes governed by norms – like the distinctive way it feels to have an occurrent feeling of obligation – should be understood to be excluded from my claim about what can be understood as a functional state.

[172] For the analogy between functionally specified mental states and poisons see Braddon-Mitchell and Jackson (2007, 44).

[173] Thus if dogs are contextually salient chocolate can be truly said to be a poison.

asymmetric dependence theory of content, but should some other theory of content prove to be superior, I would hope that what I have to say could be re-cast in terms of the alternative theory by replacing talk of asymmetric dependence with talk whatever relation that theory identifies as constituting relations of representation.

Fodor gives the example of how both horses and cows on dark nights can cause tokenings of the concept HORSE. He argues that HORSE tokens refer to horses rather than cows on dark nights because such cows' causing HORSE tokens depends upon horses causing HORSE tokens, but not vice versa. Fodor clarifies that the kind of dependence in question is actually *synchronic*. It is not simply that one's first token of HORSE was caused by horses and that cows can cause HORSE only because a horse *did*. For, Fodor suggests (1987, 109), one could have come to have a concept of HORSE from interacting only with cows that look like horses, so long as, once one has come to have the concept, the current propensity of cows to cause HORSE tokens depends upon the current propensity of horses to cause HORSE but not vice versa.

There is, I think, something of a mystery about how exactly there can be such a thing as synchronic asymmetric causal dependence. The most straightforward way of imagining a scenario where there wouldn't have been a propensity of *Y*s to cause *Z*s unless there was a propensity of *X*s to cause *Z*s (but not vice versa) is to imagine a *diachronic* asymmetric dependence relation between the propensities. That is, you imagine a scenario where *X*s started causing *Z*s, and then *Y*s came to cause *Z*s too because they had the same relevant properties as the initial *X*s. Consider, for example the following example that Fred Adams attributes to Colin Allen: "Kudu antelope eat the bark of the acacia tree. Consequently the tree emits tannin that the kudu don't like…were a human to disturb the bark of the acacia tree, it would emit tannin too."[174] As Adams reports, this is an example of an asymmetric dependence of the propensity of human disturbances to cause tannin on the propensity of antelope bites to cause tannin. But it is a diachronic asymmetric dependence – human disturbances cause the release of tannin because acacia trees evolved to release it in response to antelope bites. The trees now will simply release tannin in response to the right kind of messing with their bark. It is

[174] See Adams' page on "Fodor's Asymmetrical Causal Dependency Theory of Meaning," URL: http://host.uniroma3.it/progetti/kant/field/asd.htm. Similar remarks to those I make about the tannin release case could be made for Adams and Aizawa's (1992) original pigeon droppings case.

not as though the current propensity of human disturbances to cause tannin depends upon the current propensity of antelope bites to do so.

So in a sense the theory's restriction to relations of synchronic asymmetric dependence is a good thing, for intuitively tannin release in acacia trees does not represent antelope bites (and tannin release in response to a human disturbance is not an instance of literal misrepresentation). More generally, relations of diachronic asymmetric dependence like the forgoing are probably ubiquitous in nature – far more ubiquitous than intentional relations appear to be. To have synchronic asymmetric dependence, it seems that something like the following must be true: *Y*s currently have a propensity to cause *Z*s because (1) *Y*s resemble *X*s and (2) *X*s currently have a propensity to cause *Z*s, but the truth of (2) is not explained by (1) and the current propensity of *Y*s to cause *Z*s. It seems doubtful that these sorts of asymmetric explanatory relations obtain for *Z*s that are not mental states, which is good for avoiding the aforementioned kind of "semantic promiscuity," or predicting the existence of intentional relations where intuitively there are none. What is somewhat problematic, however, is that it is also difficult to see why or how this could happen even if *Z* is a mental state. But I shall assume that there is something about the acute, discriminating sensitivities of neural states to various patterns of inputs that enables there to be these kinds of synchronic asymmetric dependencies among the propensities of various things to cause them.

To turn back to the case at hand, we have representations of norms' prescriptions that are caused by underlying states of norm acceptance. It is as though the mind had a pattern constituted by the accepted norms, and it is checking to see if certain attitudes fit the pattern, with the representation of the norms' prescriptions playing the role of a report on whether or not a given attitude would fit. Some of these representations, we have been saying, are illusory (like appearances of commitment in cases of fallacious deductive inference) while others are veridical (like the appearance that *A* and not-*A* or *B* commit one to inferring that *B*). Now in Fodor's cases we had representations of various kinds of things, like cows and horses, as having a single property, like that of being a horse. Correspondingly we had many different kinds of things causing one sort of representation of them, and we spoke of asymmetric dependencies between the

propensity of each thing to cause the representation. In our case we have a single entity – an accepted norm – being represented as having many different properties, like prescribing this, that, or the other thing. Correspondingly, it will be most convenient to talk of a single state of norm acceptance causing many different representations of it, and to speak of asymmetric dependencies between the propensities of the state of norm acceptance to cause different such representations.

Thus, our asymmetric dependence criterion shall be that a representation of a norm's prescriptions, *R**, is illusory just in case its tendency to be generated by the underlying state of norm acceptance is synchronically asymmetrically dependent upon the tendency of other, logically distinct representations of the norm's prescriptions, *R*, to be generated by the state of norm acceptance. In our illustration from figure 8, we had a deeply accepted norm that prescribed feeling obligated not to kill sentient beings who are psychologically connected to a future good for them, which I shall abbreviate as 'connected sentients'. Now our deeply accepting this norm will have some propensity to cause representations of the norm as prescribing feelings of obligation not to kill various connected sentients, like children, psychologically typical adult humans, and companion animals. But since agents lack explicit knowledge of what the norms they accept prescribe, they may well mis-match certain other things to the pattern of connected sentients – they may take it to include embryos and people in irremediable agony, and they may, as in figure 8, simply represent the entire pattern as one that prescribes feeling obligated not to kill humans *per se*. According to the asymmetric dependence criterion, what is going on is that the propensities of the norm to give rise to the latter representations depend asymmetrically on the propensities of the norm to give rise to the former. The norm is represented as prescribing feeling obligated not to kill embryos and people in agony only because these entities resemble connected sentients. But it is not the case that the norm is represented as prescribing feeling obligated not to kill various connected sentients only because these entities resemble things like embryos and people in agony.

In essence, we might say, state *R* is a representation of state *N*'s having property *P* just in case there is a propensity of *N*'s having *P* to cause *R*, where this propensity either:

> (i) asymmetrically depends upon the propensity of *N*'s having *P* to cause distinct states of *R*'s kind (in which case *R* is illusory), or
> (ii) has asymmetrically dependent upon it various propensities of *N*'s having *P* to cause other states of *R*'s kind (in which case *R* may well be veridical, or will be so long as (i) is not also true).

But the kinds of representations that figure into the mechanism of deep norm acceptance are not just any representations of some property *P* of some other mental state *N*; they are representations of the fact that *N* is the acceptance of a norm and *P* is the property of *N*'s having as content a norm that prescribes certain attitudes in certain circumstances. Above we saw an answer to what makes a mental state *N* a state of norm acceptance and what makes another mental state *R* a state of representing the prescriptions of that norm. This was that these two states interact with each other and with certain of our attitudes (like beliefs, desires, or emotions) in the right kind of way. In particular, *N* is a state of norm acceptance for attitudes of kind *A*, and *R* is a representation of *N*'s prescriptions, just in case *N* has a propensity to cause *R*-type states and *R*-type states of a propensity to cause *A*-type states. Just as I think that these sorts of functional relations determine the kinds of states that *N* and *R* are, I think that they moreover determine the particular contents that *N* and *R* have.

More specifically, a state of norm acceptance *N* will have the content of prescribing a particular attitude *A** just in case there are states of type *R* that represent *N*, veridical ones of which have a direct propensity to cause *A** and to engender recalcitrance dissonance if this influence fails to be decisive. Veridicality here is understood in terms of asymmetric dependence as above. We thus have the following

Analysis of Contents of Deeply Accepted Norms and their Representations:
State *N* will be the acceptance of a norm that prescribes attitude *A** in circumstances *C* just in case there is some property *P* of *N* such that:

(1) N's having *P* has a propensity to cause members of the set of *R*-type states

$RT = \{ r_1, r_2, \ldots, r_N \}$,

(2) For every state r_i in *RT*, if r_i is tokened in the presence of a representation that one is in *C*, then r_i has a propensity to cause an A-type response (where A-type responses include absences of having attitudes of the relevant kind),

(3) There is a proper subset of *RT*, $\hat{R} = \{ \hat{r}_1, \hat{r}_2, \ldots, \hat{r}_N \}$, such that for every state r_i that is a member of *RT* but not a member of $\hat{R}$, there is some member of $\hat{R}$, $\hat{r}_j$, such that the propensity of *N*'s having *P* to cause r_i depends asymmetrically on the propensity of *N*'s having *P* to cause $\hat{r}_j$ [That is, *N*'s having *P* has a propensity to cause r_i because *N*'s having *P* has a propensity to cause $\hat{r}_j$, but it is not the case that *N*'s having *P* has a propensity to cause $\hat{r}_j$ because of *N*'s having *P* has a propensity to cause r_i], and

(4) For every state $\hat{r}_j$ in $\hat{R}$, if $\hat{r}_j$ is tokened in the presence of a representation that one is in *C*, then $\hat{r}_j$ has a propensity to directly cause the *A*-type response *A**, and should this influence of $\hat{r}_j$ fail to make it the case that the subject has response *A**, the subject will tend to experience recalcitrance dissonance.

These interdependent roles that together make it the case that a state *N* is a norm that prescribes attitude *A** with *R*-type states as its representations are depicted graphically in figure 9 below.

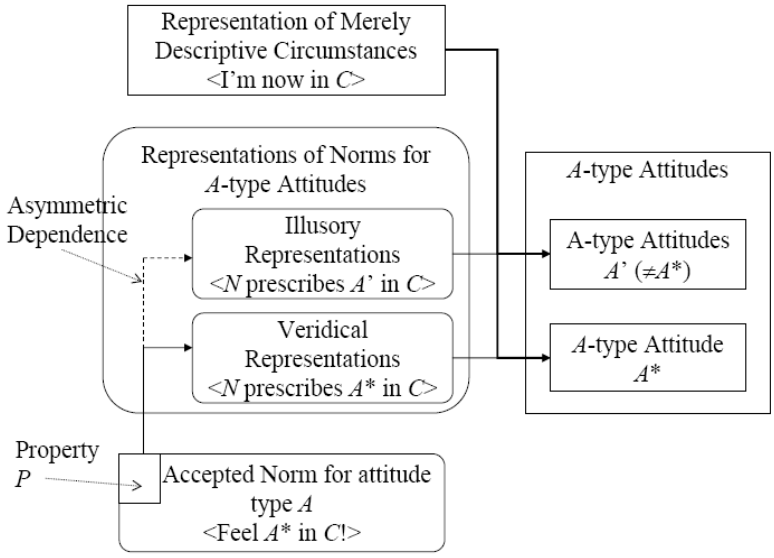

Figure 9. Content Determining Relationships Between Accepted Norms, their Representations, and the Attitudes they Govern

The sorts of representations of what our deeply accepted norms prescribe that are most directly causally sensitive to these norms would presumably be states of normative intuition. These are the un-inferred normative appearances that we solicit as raw-material for our basic methods of normative inquiry. Now if the foregoing is correct, these states will have a tendency to accurately reflect the underlying states of norm acceptance that generate them, but the tendency will be imperfect. In order to determine which intuitions are reliable, we must sift through them, seeking to determine if we can explain our having some of them solely in terms their contents bearing certain similarities to the contents of others. But if the asymmetric dependence account of representation is correct, this exercise reveals which of our intuitions accurately reflect their content and which do not. As we saw in chapter 2, intuitions, like states of perception, are distinct from beliefs or judgments – we might come to the conclusion that a given intuition is illusory but continue to have it nonetheless. But just as we can form empirical beliefs on the basis of those sensory experiences we deem veridical, we can proceed to form judgments about the prescriptions of our deeply accepted norms by relying on those intuitions we deem veridical and discounting those we deem illusory.

The checking of our attitudes against the patterns constituted by our deeply accepted norms thus has the same kind of two-level structure as our checking to see what

the external environment is like. There is an initial set of rapidly deploying representations about fit that are un-inferred and relatively incorrigible, and then there is a slower deploying set of representations that use the first as data but seek to correct and extend them through domain general processes of inference. Our judgments about what our deeply accepted norms prescribe can thus constitute an improvement over our intuitions in terms of their causal sensitivity to facts about what our deeply accepted norms actually prescribe. By relying only on intuitions that cannot be explained in terms of distinct intuitions, our judgments can be caused to conform to the facts about our deeply accepted norms responsible for veridical intuitions rather than the distortions of those facts that give rise to the asymmetry in the dependence of illusory intuitions upon the propensities of our norms to generate veridical ones.

But even our accurate intuitions might not represent the whole of the truth about what our deeply accepted norms prescribe. We may manage to get a few reliable intuitions about cases, principles, and relevant differences, but these may in and of themselves fail to deliver clear verdicts in many cases. Yet we can infer from our non-debunked intuitions about what our deeply accepted norms prescribe to a best explanation of what these norms probably are given that they have given rise to these fragments of their content. Inference to the best explanation of what our norms are given that they give rise to these accurate representations has the same epistemic properties of inferring to the best explanation of what the external world is like given that it has given rise to the perceptual states that it has. By judiciously projecting from starting points that are initially sensitive to their subject matter, inference to the best explanation has the potential to generate more general views that stay in metaphysical contact with the more general constellations of facts. For had these more general sets of facts been different, we would have had clues about this in our starting points, and we would have projected differently so as to hook onto these general sets of facts. So, anyway, is our presupposition when inferring to the best explanation in the empirical case, and so too it should be when inferring to the best explanation in the case of determining what our deeply accepted norms prescribe.

Now it seems manifestly clear that when we engage in normative thought and inquiry we do not make judgments about commitments or deeply accepted norms under

any such description. In inferring from intuitions we do not think things like "well, my deeply accepted norms appear to prescribe this and that, and here they seem reliable, so my deeply accepted norms must prescribe thus and so." The mode of presentation under which we have normative intuitions and under which we make normative judgments is, it seems, irreducibly normative. It seems, for instance that we should feel obligated to help the nearby poor, and it seems that distance cannot matter, so (we might judge) we should feel obligated to help the distant poor. But as we observed in Section 6.3, appearances or intuitions about what our norms deeply prescribe do not occur under the mode of presentation MY DEEPLY ACCEPTED NORMS SEEM TO PRESCRIBE FEELING *F* IN *C*. Rather, they might seem to occur under the mode of presentation IT SEEMS THAT I SHOULD FEEL *F* IN *C*. But as we discussed in Chapters 1 and 2, we can represent the same facts under different modes of presentation, each of which can have different functional roles in terms of their downstream consequences for other mental states. Thus, a scientist viewing stomach contractions on a computer monitor might well have visual perceptual states that represent them. But the person who simply feels hungry and doesn't know about the science of hunger might well be in mental states that represent the same facts about his stomach contractions under a different mode of presentation, namely his hunger pangs. These two different kinds of representations of stomach contractions might have very different downstream consequences. Thus the scientist may form various beliefs about the person's stomach and how it was contracting under the mode of presentation HIS STOMACH WAS CONTRACTING IN SUCH AND SO A WAY. The person who has the stomach contractions may, however, form none of these beliefs, but he will be motivated by his representations of his stomach contractions to eat food (where the scientist's representations of his stomach contractions might actually have quite the opposite effect).

In the same way, a person who is engaged in actual normative inquiry or deliberation about what to do, think, or feel would represent the prescriptions of the norms he deeply accepts under a mode of presentation quite different from, say, a super-cognitive-neuroscientist of the future who might peer at his brain states and conclude "lo, his deeply accepted norms prescribe feeling *F* in *C*." Unlike our scientist, our person who is deliberating might not token a concept with a mode of presentation like THE NORMS I DEEPLY ACCEPT at all. Rather, if the foregoing account is correct, the intuitions

of a deliberating person represent facts about her deeply accepted norms in virtue of their bearing the right nomic relations to states of norm acceptance, other representations, and the attitudes governed by the norms. They will not represent facts about her deeply accepted norms by means of any description like WHATEVER IS PRESCRIBED BY THE NORMS I ACCEPT. They will, however, be an essentially attitude-guiding mode of presentation of facts about her deeply accepted norms, for were they to cease guiding her attitudes, they would cease to play the functional roles needed to count as a representation of her norms (remember analogy to poisons). Since there is presumably no other set of mental states that simultaneously plays the roles of direct representation of deeply accepted norms and essential attitude guidance, the mode of presentation under which our intuitions represent the prescriptions of our deeply accepted norms will presumably be primitive in the sense that we cannot analyze them by means of producing synonyms for them (no matter how vague the candidate synonyms might be). This does not mean, however, that we cannot give a philosophical analysis of them, for we might still be able to say informative things about what these states are and what they represent – that is exactly what we were trying to do just above.

We might say, then, that our normative intuitions represent the prescriptions of our deeply accepted norms under a special "deliberative mode of presentation" – it is the specially attitude guiding mode of presentation that we encounter when we engage in normative inquiry or deliberation about what to do, think, and feel. Much the same can be said, I think, for the normative judgments that we make in the course of determining what to do, think, and feel. There might of course be two ways in which one could form judgments about what one's deeply accepted norms prescribe. One could take certain perceptions of what one's deeply accepted norms prescribe and then use them as evidence in forming judgments about what our deeply accepted norms actually prescribe in the way a super neuroscientist might. In doing so one would reason under the mode of presentation "my deeply accepted norms look to prescribe this and that, so they probably prescribe thus and so." But that, of course, is not what we do in deliberating about what to do. For one thing judgments about what to do, think, and feel have a direct influence on our attitudes and actions which would seem to be absent from judgments under the explicit mode of presentation of what one's deeply accepted norms prescribe. Rather, it

might seem that our normative judgments, just like our normative intuitions, are primitive or such that they resist analysis into informative synonyms. Just like normative intuitions, normative judgments would seem to get their contents from their relations to the states with which they interact – from their roles in influencing attitudes and being influenced by intuitions and ultimately accepted norms that generate those intuitions. These functional relationships are depicted below in figure 10.

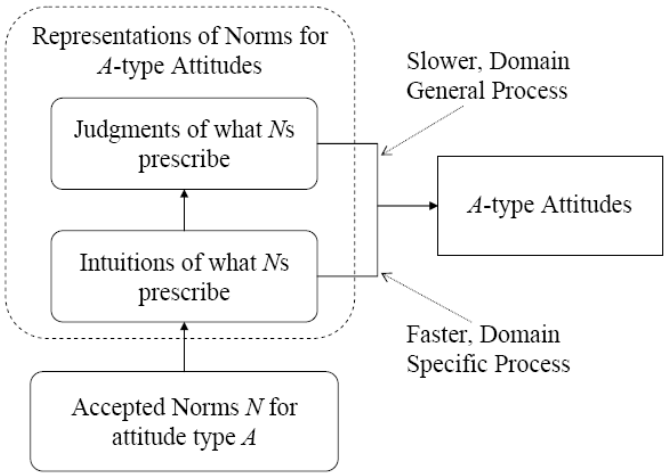

Figure 10. The Functional Roles of Judgments in Relation to Intuitions About Deeply Accepted Norms

We have seen, then, that understanding basic normative inquiry as inquiry into what is prescribed by the norms we deeply accept can explain how that inquiry guides us and how it aims at knowledge of a certain kind of fact. If basic methods of reflective equilibrium are attempts to determine the best explanation of non-debunked appearances of what our deeply accepted norms prescribe, they seek to achieve a state in which our judgments about what our norms prescribe are explained by facts about what our norms prescribe. They do this by liberating us from the illusory appearances of certain of our intuitions and extending our knowledge to cases not covered by these intuitions. It is also worth noting how this identification of normative inquiry with inquiry into the prescriptions of our deeply accepted norms can explain another feature of our reflective equilibrium methods, namely how it is a form of *a priori* access to a synthetic subject matter. What is prescribed by the norms we accept is not an analytic fact – there is nothing about the content of concepts and deductive logic alone that determines what is

prescribed by the norms we deeply accept. Yet our way of accessing these facts takes as its primary dataset intuitions about what our deeply accepted norms prescribe which we can solicit *a priori*, or independent of any sensory experience (beyond that needed to acquire the concepts in terms of which we frame our intuitions – like those of KILLING, PSYCHOLOGICAL CONNECTION, HUMAN, and so on). The determination of what best explains our non-debunked intuitions is also a largely *a priori* affair, relying as it does upon seeing which intuitions seem to asymmetrically depend on distinct other intuitions, and attempting to determine what overarching theory[175] best captures and systematizes the intuitions that are not explained away.[176]

As attractive as it might thus be to identify basic normative inquiry with inquiry into the prescriptions of the norms that we deeply accept, some philosophers have balked at identifying our more critical normative inquiries with attempts to determine our own underlying commitments.[177] The crux of the worry seems to be that we can evaluate the rationality of our own commitments, or the norms we accept, themselves. We might start out accepting norms, like *Desire only pleasure for its own sake!*, or *Feel more averse to harming members of your own race!*, that philosophical arguments convince us to reject.

It is crucial to note, however, that these philosophical arguments work by means of the same reflective equilibrium methods we have been discussing, and they seem to have the same causal and epistemic features. Coming to think that one should accept or

---

[175] Where what I mean by an "overarching theory" is simply the most general thing we can say about what we should do, think, and feel in any given situation. Such an overarching theory could in principle be quite "particularist" - it could be hedged all over the place with *ceteris paribus* clauses, and it could focus on different general factors and how they tend to interact rather than on developing any comprehensive decision procedure that mechanically spits out a complete list of prescriptions for any descriptively specified circumstance.

[176] These forms of access would rather straightforwardly count as *a priori* in the more liberal senses discussed by Boghossian and Peacocke (2000), according to which *a priori* access is access independent of sensory experience. But the access afforded by reflective equilibrium methods to the prescriptions of the norms we deeply accept may qualify as *a priori* in a stronger sense too, since it is a from of access to rather general facts that are at least in part about abstracta, and – if I am correct about what rational intuitions and insights are – it is a form of access by means of "pure reason" or cognitions of this sort alone.

It might be objected that reflective equilibrium methods are an *a posteriori* form of access to our norms because they can involve debunking arguments that pursue empirical hypotheses about the actual origins of certain intuitions. But it seems that a form of evidence (like normative intuition) can be capable of being defeated or strengthened by empirical evidence without losing its status as *a priori* in an interesting sense (see for instance (Russell 2007) and (Bonjour 1998)).

[177] See for instance (Rawls 1971, 1974) and (Daniels 1979, 1980).

reject a norm exerts direct causal influence on one's accepting or rejecting it, and the indecisiveness of this influence tends to engender recalcitrance dissonance. Deliberation about what norms to accept is a kind of *a priori* inquiry that seems capable of hooking onto a synthetic subject matter by debunking some and seeking out a best explanation of others of our intuitions about what to accept.

As such, the general considerations that support identifying the normative evaluation of a response with holding it up to norms we accept equally support identifying the normative evaluation of a norm we accept with evaluating it against higher-order norms we accept. Someone might, for instance, accept norms prescribing greater concern for her own race because they appear to be licensed by higher-order norms that prescribe greater concern for those to whom she is specially related. But this appearance may be due to a conflation of membership in the same race with somewhat correlated features like degrees of personal contact. The racist may discover that her norms about personal relations privilege only the latter by consulting her intuitions about hypothetical cases and the relevance of what race-membership actually comes to.

To illustrate how this kind of influence of higher order norms on lower order norms and attitudes might work, suppose that Bugsy is your friend of a different race than you and Freddy is a stranger of the same race as you. Suppose further that you deeply accept a higher-order norm for considerations of special relationships N_H. In fact N_H prescribes feeling more strongly obligated to help someone the greater your degree of personal contact with her, but as is common with deeply accepted norms you do not know this. You start off mistaking the relevance of the degree of personal contact for the relevance of a feature that is often correlated with the degree of personal contact: namely the degree of blood-relation. As a result you (falsely) represent N_H as prescribing the acceptance of a first order norm N_R, which prescribes feeling more strongly obligated to help members of your own race. This mistaken representation of N_H's prescribing accepting N_R causes you to accept N_R. You then (accurately) represent that N_R prescribes feeling obligated to help Freddy over Bugsy, and as a result you feel obligated to do so.

But suppose that after doing your basic normative inquiry you tease apart blood relation and degree of personal contact (you thought, for instance, about adopted children, about how to respond if people you took to be of the same race turned out –

unbeknownst to you and them – to have descended from aliens or to have been the product of lightning hitting a swamp, or whatever). As a result you cease to represent N_H as prescribing accepting N_R, you now accurately perceive that N_H in fact prescribes accepting a different lower order norm, namely N_F, which prescribes feeling more strongly obligated to help your friends than strangers (regardless of the race of either). This veridical representation of N_H's prescribing N_F causes you to cease accepting N_R and to come to accept N_F instead. You then (accurately) represent that N_F prescribes feeling obligated to help Bugsy over Freddy, and as a result you feel obligated to do so. These influences of N_H on your acceptance of N_R and N_F, and corresponding feelings of obligation, are depicted in figure 11 below.

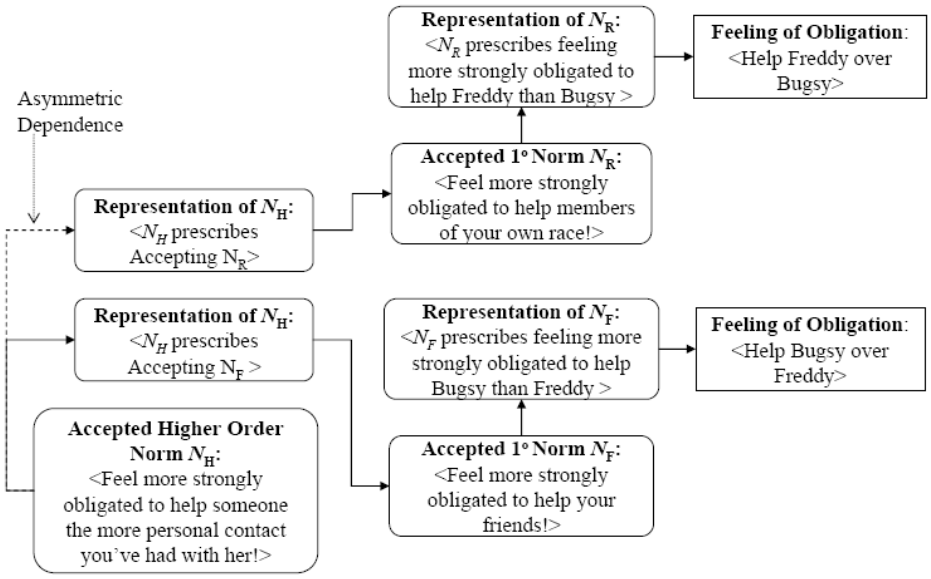

Figure 11. The Influence of Higher Order Norms on the Acceptance of Lower Order Norms and Attitudes

To accept a norm that prescribes response *A* is roughly to be in a state which is such that representations of it cause responses of *A*'s kind on pain of recalcitrance dissonance, and accurate representations of it cause response *A*. To accept a "first-order" norm for belief or motivation is to be in a state that in this way regulates beliefs and motives without regulating any intermediary states of norm acceptance. To accept an "n+1st-order" norm ($n \geq 1$) is to be in a state that regulates the acceptance of n^{th}-order norms. By regulating the acceptance of lower-order norms, the higher-order norms we accept actually govern responses of the kind prescribed by these lower order norms, and

constitute norms for these responses as well.[178] We can, for instance, count the prescription to feel more strongly obligated to help Bugsy over Freddy as a prescription of N_H itself, which it mandates through mandating the acceptance of N_F which prescribes this attitude. What we have seen, then, is that the causal and epistemic features of deliberation about whether to accept and have the responses prescribed by lower-order norms can be explained by identifying it with inquiry into whether the lower-order norms' prescriptions are seconded by higher-order norms. Judging that one should respond as a lower-order norm prescribes thus seems contingent upon an appearance that the response is ultimately prescribed by higher-order norms that endorse the lower-order norm.

Now, at some point our psychologies will run out of norms against which to assess other norms.[179] We will terminate at some highest-order or most fundamental norms that govern our acceptance or rejection of all lower order norms and their prescriptions. Since norms have prescriptions for the responses they regulate, the fundamental norms that regulate our acceptance of norms for belief and motivation will constitute our most fundamental norms for beliefs, motives, and the actions they motivate.

If this account of normative inquiry is correct, we are in no position to say exactly what our fundamental norms are until we know the true general theory of what to believe and pursue. I suspect that our most fundamental norms for belief include deductive norms like *modus ponens* and ampliative norms like inference to the best explanation.[180] Our coming to accept these norms may have been an evolutionary adaptation that enabled our ancestors to form accurate beliefs under a wide range of novel and complex conditions.[181] Similar selection pressures may have caused us to accept fundamental norms for motivation that enabled our various motivational systems to play their adaptive roles more flexibly and in ways better suited to novel and complicated environments. I

---

[178] Thus, to keep n^{th}-order norms from themselves counting as n-k^{th}-order norms (for n-$k > 1$), we should understand n+1^{st}-order norm acceptance as a state that regulates n^{th}-order norm acceptance and the acceptance norms of no higher orders.

[179] Which will be the case, moreover, for any entity with finite psychological capacities.

[180] Where determining what it is for something to be a "best explanation" is a serious task of normative epistemology.

[181] For an excellent and engaging discussion of these environmental conditions, see (Quartz and Sejnowski 2002).

suspect that these norms are rather complex. But proponents of different theories about what we have reason to do will have their own views about these fundamental norms. A utilitarian about rationality might, for instance, contend that we all fundamentally accept the principle of utility and only tend to judge that utilitarianism is false because we mistake things that are typically optimific for things that our norms prescribe doing for their own sakes.

6.6. Norm Descriptivism vs. Norm Relativism

If our inquiries into what to do, think, and feel can be explained as inquiries into what the most fundamental norms we deeply accept prescribe, it might seem that questions and judgments about what to do are nothing other than a particular mode of presentation of questions and judgments about what these most fundamental norms prescribe. This would be the mode of presentation of our most fundamental norms that we encounter in deliberation, in virtue of these states bearing the right nomic relation to the states that constitute our acceptance of these norms. One might, however, attempt to resist the inference from the foregoing explanation of normative inquiry to the identification of normative thought with thought about the prescriptions of norms we deeply accept. One might do this if one thought that our normative judgments "point through" the norms we accept to something else. Much as representations of magnified images can be about the microscopic objects that caused them, the idea might be that the normative judgments are about something in the external world that caused us to accept the norms we do.

The problem with this line of resistance is that the causal genesis of our fundamental norms seems irrelevant to the content of our normative judgments. As we saw in Chapter 1, considerations of parsimony dictate that no *sui generis* normative facts explain our acceptance of norms. Suppose it turned out that we accept the fundamental norms we do because of some particular evolutionary story, or because of a certain divine creation, or because of the details of how we sprang into existence a few moments ago as a result of lighting hitting a swamp. None of this seems to make a difference to what we are thinking about when we make judgments about what to do, think, or feel. Facts about

the genesis of certain norms can matter to our basic inquiry into whether to accept them. But when this happens, we take these facts about origin to bear on whether the norms fall short of an independent standard that guides our acceptance or rejection. Since our fundamental norms are the ultimate standards by which we assess our norms, their origins cannot matter in this way. The normative judgments that guide us seem to point as far as the prescriptions of our fundamental norms but no further.

There, might, however, be different ways of thinking about what it would be to identify judgments about what to do, think, and feel with judgments about the prescriptions of the most fundamental norms one accepts. The main ways in which they differ concern how we should think about judgments concerning what people other than oneself should do, think, and feel. One way of thinking about this question is the following. When we make judgments about what another agent should do, think, or feel, we make judgments about the same things she does. The judgment we express with 'you shouldn't ϕ' contradicts her judgment that she should. The agent seems able to assess the truth of our judgment using the same reflective equilibrium methods that she uses to assess her own, and coming to believe that our judgment is true will directly guide her responses in the same way as her own. Our thoughts about her reasons thus seem to be thoughts to the effect that the answers she seeks in deliberation are thus and so.

If this is right, then the considerations in favor of the identification of thought about what to do, think, and feel with thought about what one's fundamental norms prescribe suggest that our judgments about another agent's reasons are judgments about what her most fundamental norms prescribe under a mode of presentation that we fix by reference to her deliberations. This yields the following general view of judgments about what an agent should do, think, or feel:

Norm Descriptivism:

To judge that agent *A* should ϕ (where ϕ can be any response – believing, feeling, intending, or doing something) is to believe under a mode of presentation derived from *A*'s deliberations that the most fundamental norms that *A* deeply accepts prescribe that *A* ϕ.

Just like Norm Expressivism, Norm Descriptivism is a general theory of normative thought, which offers us among other things an account of what it is to judge that it is fitting for an agent to have an attitude. It is to represent under the relevant mode of presentation that the most fundamental norms she deeply accepts prescribe that she have it. If you are the agent whose responses you are assessing for fittingness, this is exactly the kind of judgment that will cause you to have a direct propensity to have the attitude. If someone else is the agent whose responses you are assessing, her coming to believe the content of what you judge under the same mode of presentation (i.e. that derived from her deliberations) will have a direct propensity to cause her to have the attitude. In this way Norm Descriptivism offers us a way to explain the causal and epistemic features of fittingness assessments without invoking the concepts that the fitting attitude analyst is trying to analyze. So Norm Descriptivism can work equally well to solve the WKR problem of giving an independent characterization of fittingness judgments, which can explain not only how they guide us but how ethical inquiry aims at making contact with ethical facts understood as facts about which valenced attitudes are prescribed by the most fundamental norms we deeply accept.

One might worry that Norm Descriptivism fails to capture the apparent platitude that to judge that another agent has reason to do something is to endorse her doing it, or to think that one would have reason to do it oneself if one were in her circumstances. Of course, Norm Descriptivism easily captures this platitude if "an agent's circumstances" are allowed to include facts about the norms she most fundamentally accepts. But the kind of circumstances intended by the platitude are presumably those which authoritative norms have prescriptions for rather than those which determine which norms are authoritative. The alleged platitude thus seems to amount to the idea that that the same basic set of norms is authoritative for each agent, or prescribes what each agent should in fact do.

An alternative way of putting the alleged platitude about endorsement might be to say that a feature of a situation that is a normative reason for one agent to respond in a certain way must be a normative reason for any agent to respond in that way if she too is

in a situation with that feature. Of course, if Dancy (2004) is right, then some considerations might be completely turned on or off as reasons due to the presence of other considerations – strengtheners and weakeners – which do not themselves count in favor of responding in a certain way but which increase or decrease the strength with which other considerations do so (and in the limit entirely enable or disable the ability of other considerations to favor responses at all). But presumably we can take account of this; we could understand the idea to be that if a consideration *R* is a normative reason for one agent but not a reason for another who faces *R*, then there must be some consideration *Q* which one of the agents faces and the other does not, such that *Q* is a strenghener or weakener for that agent and would operate in the same way as a strenghener or weakener for the other agent were she to face *Q* as well.

Now one thing that I should clarify about Norm Descriptivism is that it does not say that the fact that an agent accepts the fundamental norms she does is itself a reason (or strengthener or weakener of a reason) for her to respond in any way. For to figure out what is a reason, strengthener, or weakener one turns to one's basic methods of normative inquiry, and Norm Descriptivism is trying to give an account of what goes on when that happens. According to Norm Descriptivism, to think that a feature of a situation is a normative reason for an agent to respond in a certain way is to represent it under the appropriate mode of presentation as contributing to its being the case that the fundamental norms the agent accepts prescribe that she respond in that way. Norm Descriptivism thus gives us the following analysis of thoughts about normative reasons:

> **Norm Descriptivist Theory of Judgments About Normative Reasons**:
> To judge that *R* is a normative reason for agent *A* to ϕ is to believe under a mode of presentation derived from *A*'s deliberations that *R* contributes to making it the case that the most fundamental norms that *A* deeply accepts prescribe that *A* ϕ.[182]

[182] Similarly for strengtheners and weakeners:

> To judge that *S* is a strengthener of a normative reason *R* for agent *A* to ϕ is to believe under a mode of presentation derived from *A*'s deliberations that *S* increases the extent to which *R* contributes to making it the case that the most fundamental norms that *A* deeply accepts prescribe that *A* ϕ.

Since the fact that I most fundamentally accept a given norm does not contribute to making it the case that that norm prescribes anything (unless perhaps it is a really bizarre norm), it will according to Norm Descriptivism be false (and silly) to judge that it is a reason for me to respond in any way. If I accept fundamental norms that *ceteris paribus* prescribe believing theories with the fewest numbers of kinds of entities, then the fact that a theory posits fewer kinds of entities than its rivals contributes to its being the case that my fundamental norms prescribe that I believe it. If I accept fundamental norms that ceteris paribus prescribe feeling obligated not to kill connected sentients, then the fact that Bugsy is a connected sentient contributes to its being the case that my fundamental norms prescribe feeling obligated not to kill Bugsy. That a theory posits fewer kinds of entities and that Bugsy is a connected sentient are plausibly regarded as ultimate reasons - they count in favor of believing the theory and feeling obligated not to kill Bugsy independent of our having reason to believe or feel anything else that they are ways of believing and feeling. Of course, it is not self-evident that these are ultimate reasons; one needs one's reflective equilibrium methods to determine that they are. But the epistemic question of how we can tell that something is a normative reason is different from the metaphysical question of whether it is only a reason because it is a way of doing, thinking, or feeling something else that we have reason to do, think, or feel.

It would be bizarrely self-absorbed to say that the reason for you to believe a theory or feel obligated not to kill people is that your norms prescribe thus and so. Surely my reasons to believe theories and feel obligated not to kill people are things about those theories and people rather than things about my head. And this is exactly what Norm Descriptivism gives us. For Norm Descriptivism seeks to explain the status of certain

---

> To judge that *W* is a weakener of a normative reason *R* for agent *A* to ϕ is to believe under a mode of presentation derived from *A*'s deliberations that *W* decreases the extent to which *R* contributes to making it the case that the most fundamental norms that *A* deeply accepts prescribe that *A* ϕ.

Dancy (2004) criticized attempts to understand the notion of a reason in terms of the notion of what one should do overall. Since the notion of what an agent's norms prescribe is an overall notion, these are essentially attempts of that kind. But by appealing to the metaphysical notion of something's helping to make it the case that the truthmaker for an overall verdict obtains, I believe that these kinds of accounts can avoid Dancy's worries. I think that contributing to making something the case should be a perfectly familiar notion – adding a stone to one side of a scale contributes to making it the case that the scale will tip in favor of that side.

considerations (perhaps like the foregoing) as ultimate reasons for an agent to respond in a certain way in terms of their contributing ultimately to making it the that that agent's fundamental norms have the relevant prescriptions. Facts about the fundamental norms you deeply accept simply explain what makes it the case that you have the ultimate reasons you do, and how you can find them out by engaging in your basic methods of normative inquiry.

Now Norm Descriptivism will predict that the same norms are authoritative for all agents, or that all agents have the same reasons, just to the extent that all agents deeply accept the same fundamental norms. I think, however, that our experience with shared normative inquiry suggests that this is largely the case. A shared set of fundamental norms seems to me to be the best explanation of the similarity we see among our intuitions, what we take to be problems for these intuitions, and what we take to be problems and prospects for theories that seek to unify them. There is far more similarity in people's *intuitions* about cases, relevant differences, and principles (once fully clarified and understood) than there is among their views about how to account for them. Even where specific intuitions differ, there seems to be substantial similarity in surrounding intuitions, including about such epistemic considerations as what kinds of origins of intuitions look suspicious and what kinds of features look like theoretical problems for intuitions and theories that seek to account for them.

Moreover, this kind of overlap in intuitions and the problems and prospects of accounting for them is what seems to make shared normative inquiry possible in the first place. Against a background of shared fundamental norms that generate such overlap, we can speak simply of 'the thing to do in a circumstance', which will be the same for all of us. Perhaps, then, we tend to think that the same norms are authoritative for everyone because our shared normative inquiries presuppose that we all accept the same fundamental norms, and this presupposition is pretty much correct. If this is right, then Norm Descriptivism and the fact that we all accept the same fundamental norms can explain how the same norms are authoritative for everyone, or how everyone has the same reasons if faced with the same circumstances, and consequently capture the platitude about endorsement.

As much as it might seem plausible as a way of accounting for our shared normative intuitions and dispositions to perceive problems for these intuitions and for normative theories, it may be worth asking why it would be that we all deeply accept the same fundamental norms. I mentioned earlier that I suspect that we came to accept the fundamental norms we do as a universal human adaptation to variable and complex environments. This might be plausible enough when it comes to epistemic norms, or the norms we deeply accept for belief. Believing theories that are *ceteris paribus* the simplest would (we think) enable organisms to believe theories that would on the largest number of occasions true, and having true beliefs would tend to confer selective advantages. It is unlikely, however, that every detail of our epistemic norms would have been an adaptation. If we (as our reflective equilibrium methods may reveal) accept norms that prescribe believing in fewer kinds of entities but not fewer entities, it is unlikely that this was singled out by evolution as exactly the right epistemic norm to have. Of course, we believe (given how we represent the prescriptions of our norms) that theories that posit fewer kinds of entities should be expected to be true more often than theories that posit fewer entities. But it seems doubtful that evolution would have been sensitive to the difference. More likely, evolutionary processes just (as it were) "programmed in" some rules that could guide theory construction in a way that was often enough sufficiently accurate subject to the many constraints on what kind of neural tinkering was feasible. These rules just happened to be those that went for small numbers of kinds of entities rather than small numbers of entities. That it was numbers of kinds rather than numbers of entities might well just have been an evolutionary byproduct, or what people following Gould and Lewontin (1979) have come to call "spandrels" – features that look to us that they might have been adaptive but that really just came about as a side effect of something else that was.

It is important, then, not to look to the evolutionary history of our norms for some further justification of their prescriptions. The considerations spoken of by our fundamental norms will be normative bedrock, and the reflective equilibrium methods we use to determine what these norms prescribe will be the bedrock of normative epistemology. We cannot help but regard ourselves as having in a sense gotten lucky in

terms of what epistemic norms evolution programmed in. If we accept norms that prescribe believing in fewer kinds of entities (and accurately represent this fact) we will think that evolution gave us the norms that are best at tracking the truth. But if evolution had given us norms that prescribe believing in fewer entities we would have thought that those were the most reliable and counted our selves lucky for having been given them instead. The spirit in which we are looking to evolutionary considerations now is simply to understand why it would be that we all deeply accept the same fundamental norms. Why, that is, what is normative bedrock for me is normative bedrock for you, not whether we can dig deeper than normative bedrock.

Of course, if normative bedrock is the same for you as it is for me, it is notorious that we often don't know what that bedrock is and that we often disagree on what it looks like. So, we need to ask, why would evolution have given us norms with prescriptions that we can't figure out? If even 21st century philosophers, with thousands of years of research to draw on and their entire lives to spend asking what the norms prescribe cannot figure out their prescriptions, what use would they have been to your typical ancestral hominid who was supposed to get guided by them in a fitness enhancing way? The answer, I think, has to do the fact that the exact nature of the prescriptions of our norms is an evolutionary byproduct. We needed some sort of pattern against which we could check our attitudes if they were to play their adaptive roles more flexibly, and it had to be the kind of pattern that it would have been adaptive to approximate if you looked only so far into what it seemed to prescribe. But since no one was going to look at the pattern all that closely, there was no need to make it adaptive in every detail, or to make it the kind of thing that it would have been adaptive to conform to at all if you really knew what it was. Like most evolutionary processes, the shaping of the pattern itself would presumably have been an entirely haphazard affair. Various mutations in genes coding for the neural states that constitute our acceptance of norms would have thrown in a tendency to respond to one set of features in one way and a tendency to respond to another set of features in another, and out of the soup would have emerged the thing that we are actually all trying to approximate. There is absolutely no reason to think that what emerged would have had any clear adaptive rationale apart from getting good enough guidance from it if you skimmed the surface of what it prescribed.

While it wouldn't have mattered exactly what deep pattern we were all trying to approximate, it would, I think, have mattered that it was pretty much the same for everyone. For the environment of adaptation that likely gave rise to our systems of normative governance was complex in part because of the intensive forms of social interaction it involved. As we saw in Chapter 5, it seems that the adaptive function of the normative governance of moral and honor system emotions had much to do with flexibly facilitating social cooperation. Now even slight differences in deep normative would probably give rise to deep differences in intuition, which could lead quickly to irreconcilable differences in normative outlook. If what you adduce as reasons to feel obligated simply cut no ice with me, and we are unable to get anywhere in about five minutes of unfocused conversation, then we probably won't be able to coordinate very well if we are ancestral hominids. Of course, if anything like the picture I have been painting of basic normative inquiry is correct, we have to recognize that social forces can swing our beliefs about what our deeply accepted norms prescribe very, very far off base from the facts about what they actually prescribe. But if we look more closely there seems to be a method to the social madness. There are standard kinds of conflations and standard kinds of intuitive tendencies that distortionary social mechanisms exploit. Extreme racism wouldn't be able to get very far if people didn't have some kind of tendency to have intuitions about special obligations, or to have intuitions about the wrongness of certain acts (whether fabricated or genuine) for them to then confusedly project onto all members of the despised race. Social brainwashing seems to need a shared base of intuition-tendencies to work with, and that would be hard to secure in the absence of a shared base of deeply accepted norms.

Of course, if it is a fact that all agents deeply accept the same fundamental norms, it seems to be a contingent fact, and unlikely to be a fact at all if we take its universal quantification to be literally unrestricted. Mutations, quirky developmental trajectories, and injuries make it almost inevitable that some humans will fail to have traits that were universal human adaptations. One could try saying things like "we rigidly fix the reference of 'agent' as *being that accepts the fundamental norms we actually do!*" But this merely distracts attention from more interesting questions about what correctly

answers the attitude and action guiding questions asked by beings that seem able to figure out such answers by means of reflective equilibrium methods.

If it is true that inquiry into what to do, think, and feel is inquiry into the prescriptions of one's most fundamental norms, then a respect in which a being's fundamental norms differ from ours is a respect in which her deliberations about what to do, think, and feel aim at something different from ours. It is a respect in which perfectly careful inquiry from intuitions that accurately represent their subject matter would lead each of us to think that we should have different responses. It seems quite natural to think that in such cases both of us would be correct, and that we simply have reason to respond in different ways. Of course, if the other agent has reason to do something that we have serious reason to prevent, we might have very good reason not to let her know this, and even to lie to her and to trick her about her reasons if need be. But conflicts of interest and reasons not to let an agent think she has reason to do something no more signal the presence of disagreement in normative judgment here than they do in competitive games.

Some people might, however, accept my argument for identifying normative inquiry with inquiry into the prescriptions of one's own most fundamental norms but think for some reason that it is very important for us to be able to truly or sincerely say of agents that accept different fundamental norms that they have reason to do what we have reason to do. These people might be reluctant to abandon pre-theoretical intuitions to the effect that it is necessarily the case that, come what may, we must have the same reasons if faced with the same circumstances. To retain these intuitions, they might try to reduce our judgments about what other agents should do to judgments from our own deliberative perspective about what to do in their circumstances.

These opponents of Norm Descriptivism may also think that there is far less similarity than I do in the fundamental norms that people deeply accept, and find themselves reluctant to think that the savvy participant in normative discussions must go in for so much trickery. If so, they will need an account of how sincere expressions of normative judgments made from deliberative perspectives with different targets can seem like genuine disagreements rather than mere differences in attitude. We saw the idea of

an account somewhat like this in our above discussion of Norm Expressivism. In the current context the idea would be that there is a psychic mechanism by means of which simply expressing our judgments about what our fundamental norms prescribe tends to cause our audience to accept these norms as well. This would support the following kind of Norm Relativism:

> **Norm Relativism**:
> To judge that agent *A* should ϕ is to believe under a deliberative mode of presentation that the most fundamental norms that one deeply accepts oneself prescribe ϕ-ing in *A*'s circumstances.[183]

A first problem with this view is that its attempt to retain our ability to truly or sincerely judge that agents with different fundamental norms should do what we should do ends up portraying all normative discussion as dishonest and akin to brainwashing. Telling another agent that she should do something purports to tell her something that, just like her own judgments about reasons, is true just in case it correctly answers her deliberative questions about what to do. But on the relativist account one's judgments about her reasons are not what they purport to be - their truth is in no way dependent upon their ability to correctly answer the agent's deliberative questions about what her fundamental norms prescribe. Like brainwashing, the influence of interpersonal normative discussion on an agent's attitudes and behavior would be independent of the influence exerted by her reasoning about what to do and the facts that her reasoning seeks out.

A second problem with this kind of view is that, as Egan (2006) has pointed out, it entails that each agent can arrogantly claim to be immune to a kind of normative error to which others are prone. This kind of Norm Relativism subscribes to my account of access to one's own reasons in terms of the identity of the deliberative question about

[183] It should be noted that there could be a version of Norm Expressivism that took on board the deep model of norm acceptance and identified judgments about how other agents should respond with (1) the acceptance of a set of fundamental norms, and (2) the belief under a deliberative mode of presentation that it has such and so prescriptions for the other agent's circumstances. Unlike the former version of Norm Expressivism, this version would not be too importantly different from Norm Relativism, and would face the same problems outlined in this section.

what to do with the question of what one's fundamental norms prescribe. It thus entails that if one manages to accurately discern the prescriptions of one's most fundamental norms, one is guaranteed to be correct about what one has reason to do. Norm Relativism also holds that you should think that others are correct about what they should do only insofar as what they think they should do is what is prescribed by your own fundamental norms. Since other agents may accept fundamental norms that prescribe doing otherwise, Norm Relativism entails that they might successfully discern what their fundamental norms prescribe and yet be wrong about what they should do.

Thus, according to Norm Relativism, you should think that other agents can deliberate ever so carefully from intuitions that accurately track their content, yet fail to get things right about what to do. You should thus think that they are vulnerable to a kind of "normative blindness" – their fundamental norms and the intuitions that reflect them might simply fail to correspond to something like an independent normative reality. But since the standards of this independent normative reality are set by your own fundamental norms, it is inconceivable that you could suffer from the same fundamental blindness as other agents. Surely this seems wrong.

I think that these costs of attempting to retain intuitions that agents with different fundamental norms have the same reasons we do outweigh their benefits. They seem to run afoul of the facts that the dynamics of normative discussion mirror those of first-person deliberation and that other agents have the same kind of access to their reasons that we do to ours. Norm Descriptivism's identification of judgments about another agent's reasons with judgments about the target of her deliberations seems to provide a far better explanation of these features of normative thought. Moreover, there is reason to think that interpersonal normative inquiry presupposes a background of shared fundamental norms, and the general truth of this presupposition is suggested by our experience with normative inquiry and evolutionary considerations. If this is right, it would not be surprising for us to mistake reasons shared by almost all human agents for reasons shared by all conceivable agents.

6.7. The Moorean Challenge

I have thus argued that an analysis of normative judgments in terms of deeply accepted norms can make best sense of basic normative inquiry, and that Norm Descriptivism is the best such identification on the grounds that it can explain how we are thinking about the same thing as another agent when we make judgments about what she should do, think, and feel. But one might object to Norm Descriptivism on grounds reminiscent of Moore's "open question argument." It looks like Norm Descriptivism gives us an analytic identity between facts about what one should do and facts about the prescriptions of one's fundamental norms. But it certainly seems that one can ask, "Should I really continue to accept and follow the prescriptions of my fundamental norms?" How does Norm Descriptivism account for this?

If one's question about whether to accept and follow one's fundamental norms is a genuine question about what to accept and do, then answers to it must directly guide what one accepts and does in the ways characteristic of normative judgments. It thus seems that one cannot be asking such things as whether continuing to accept the fundamental norms one does will make one happy, satisfy one's desires, be approved of in one's society, or conform to rules that one does not currently accept.

Expressivists would contend that such questions about what to accept are not about anything descriptive at all. But for reasons I have reviewed, I find it rather obscure what they would then be. Perhaps they are supposed to express requests for or signals of receptiveness to the kind of influence that relativists and expressivists might posit to make sense of normative discussion. But questions about what to accept do not request any old kind of influential utterances; they request correct answers to what they are asking. The search for such correct answers seems to aim at a state of knowledge in which the truth of one's beliefs about the answers helps explain why one holds them. But this kind of explanatory work can only be done by descriptive facts.

I think that the only way to explain at once the causal and the epistemic features of questions about whether to accept and follow one's fundamental norms is to identify them with evaluations of these norms against themselves. This is what Norm Descriptivism does by interpreting it as a question of whether one's fundamental norms,

conceived under a deliberative mode of presentation, prescribe accepting these norms, conceived of as THE FUNDAMENTAL NORMS I ACCEPT. Since the deliberative mode of presentation does not refer to one's norms under the latter description, it is by no means obvious to the agent that her question is an assessment of her fundamental norms in terms of themselves.[184]

In fact, it seems possible to conceive of agents who accept fundamental norms that prescribe rejecting themselves or their own prescriptions. In wondering whether one should accept and follow the prescriptions of one's fundamental norms, one can be said to be wondering whether one's fundamental norms are self-undermining in this kind of way. Since there seems to be no indication from our reflective equilibrium methods that we are committed to self-undermining principles with no deeper, non-self-undermining principles to fall back on, we seem to have little to fear from this possibility. But one's acceptance of self-undermining fundamental norms is still quite a genuine possibility to be concerned about and to wonder about *de re* when one wonders whether to accept and follow one's own fundamental norms.

A second concern about Norm Descriptivism might be a third-person analogue of the foregoing. It might still seem at least coherent to think that another agent's fundamental norms are simply irrational or crazy, or that she has reason not to do certain things, like torture innocents just for fun, whether or not her fundamental norms prescribe against them. I think, however, that these appearances of coherence stem from a lack of appreciation of what distinguishes an agent's having reason to do something from other normative phenomena. Return to the distinction that we emphasized in chapter 3's discussion of Humeanism between:

---

[184] Although Norm Descriptivism identifies normative judgments with representations of agents' fundamental norms under modes of presentation that do not describe them, it still maintains an analytic identity between facts about agents' reasons and facts their norms. According to the view, it is an analytic truth that a state would not count as a normative judgment if it failed to play the right kind of roles in deliberation and attitude guidance, which involve representing one's fundamental norms under a deliberative mode of presentation (by bearing the right kind of nomic relation to states of norm acceptance).

I believe that a similar kind of analytic identity is maintained by analytic functionalists about qualia who think that our ordinary qualia concepts are phenomenal. According to these views, it's analytic that whatever states play the qualia-roles are qualia, even though we ordinarily represent these states with phenomenal concepts that do not describe the qualia-roles. The analyticity is maintained by the contention that it is an analytic truth that a state that failed to bear the right kind of representational relation to whatever states play the qualia roles would not count as a (phenomenal) qualia concept.

(a) an entity's doing something the occurrence of which is bad, or something we should hope it doesn't do and oppose if we can, and
(b) an entity's doing something that it has reason not to do.

The occurrence of natural disasters and attacks by sharks and coyotes are instances of (a) but not (b), while the sub-optimal play of our opponents in fair competitions are instances of (b) but not (a).

As the Humean insisted, we must ask: why, in any given case, should we think that someone's doing something like torturing innocents for fun is not only something we should hope she doesn't do and prevent her from doing, but moreover something she has reason not to do? The Humean's answer, which in chapter 3 we accepted and used against her, was that in thinking that an agent has reason to do something, we do not simply think that we have reason to oppose her doing it; we think moreover that she could correctly reason her way to refraining from doing it. Again, correct reasoning is not just any causal process whereby entities come to do what we want them to do or what we might think we should do if we (with our psychology) were in their circumstances. Neurally altering a shark so that he no longer attacks innocents does not make it the case that he has correctly reasoned his way to refraining from doing so.

In chapter 3 we saw that correct practical reasoning extends beyond instrumental reasoning. It includes basic normative inquiry pursued by our reflective equilibrium methods. But what we have seen here is that those reflective equilibrium methods seem to be best explained as attempts to determine what is prescribed by the most fundamental norms we accept. Nothing else seems to have at once the right causal and the right epistemic features of basic normative inquiry into what to do. If this is right, then we can only correctly reason our way to doing what our most fundamental norms in fact prescribe we do. But if this is right, how could we maintain that an agent has reason to do something even though her fundamental norms do not prescribe doing it?

I believe that understanding judgments about someone's normative reasons as judgments about what she could correctly reason her way to doing, thinking, or feeling is a directly credible way of making sense of the substance of these judgments and how

they differ from other normative assessments (like those about how we should hope she responds). But the dependence of reasons on the possibility of this kind of correct reasoning seems important for explaining two further things about the domain of things to which normative reasons apply. First, it seems important for explaining why it is genuinely incoherent to think that reasons apply to the responses of entities like volcanoes, infants, and sharks on the assumption that these entities are as we take them to be. If someone were to think simultaneously that volcanoes are as we take them to be (with no mental life at all) and yet that they have reason not to, say, erupt, he would seem rather straightforwardly incoherent.

To be sure, beings like infants and sharks are unlike volcanoes in that they are capable of well being or welfare; of things literally going better or worse for them. But a response's being good or bad for a being is distinct from her having reason to have it. As Darwall (2002) has convincingly argued, judgments about a being's welfare are judgments about our reasons to want things for her out of care for her, which in no way requires her to be subject to reasons herself. Darwall draws on the exact same kinds of considerations that we seen in favor of fitting attitude analyses of other ethical concepts to support an analysis of the notion of a being's welfare in terms of what it would be fitting to want for her sake out of fitting care or sympathetic concern for her. Thus Darwall argues:

> …it seems possible for two people who care about someone, *S*, to coherently disagree about whether something, *X*, is good for *S*, even though they agree completely about all the non-normative facts concerning *X* and *S*…
>
> Suppose, for example, that *X* is a pleasant illusory belief of *S*'s, say, that *S*'s novel has sold 10,000 copies (when in fact it has sold only 12). It would seem that the two people could be agreed about everything else, but simply disagree about whether this pleasant illusory belief is good for *S* or makes some contribution to his welfare, other things being equal. In such a case, it is hard to see what else they could be disagreeing about other than whether *X* is to be (ought to be) desired for *S*'s sake, or, equivalently, whether it would be rational (warranted, justified, make sense) for someone who cared about *S* to desire *X* for *S* [or as Darwall explains on p.7-8, whether there obtain reasons to desire *X* that are conditional "on a hypothesis…that the cared for is *worth* caring for"] (Darwall 2002, 11).

What would best explain what is common to all coherent judgments that can take such widely different things to be good for a being, then, is that they are judgments that, if it is

in fact fitting to care for or feel sympathetic concern towards the being, then it is fitting to desire the thing out of this care for her. Given the general way in which judgments of the fittingness of attitudes directly guide them, this would explain the intimate connection that judgments about a being's welfare have to care, desire, and action. Judging something bad for a being need not always exert direct causal pressure in the direction of being aversed to it out of care for her, for one might think that the being is a culpable agent who deserves punishment (where one's reasons to care for her, perhaps mild though existent, are far from one's mind). But they do seem to exhibit the following conditional structure: to the extent one thinks one should care about the being, they exert direct influence on one's being aversed to it out of care for her.

To think beings like infants and sharks not only capable of welfare but moreover subject to reasons, we would seem to have to think them candidates for the kind of rational criticism involved in calling someone an idiot for making a foolish decision. Infants and sharks are of course capable of greater or lesser intelligence or learning ability, but it seems incoherent to hold them genuinely rationally criticizable on the assumption that their minds are as we take them to be. It is true that rational criticizablility involves more than simply failing to respond to the reasons one has; an agent can fail to do this but be rationally exculpated on account of diminished responsibility. Indeed, we have seen reasons in chapter 4 to think that all agents will be morally exculpated, and this may well mean that they will be rationally exculpated as well. But the way beings like infants and sharks lack rational responsibility for their attitudes and behavior still seems importantly unlike that of, say, an otherwise psychologically typical adult human whose perennially recalcitrant emotions always sway her against her better judgment. Infants and sharks are incapable of this very kind of "better judgment" that has the peculiar causal and epistemic properties of judgments about reasons. But it seems that one must be able to correctly reason one's way to these kinds of judgments in order to be subject to reasons at all.

That one can only have reason to do what one can correctly reason one's way to doing also seems important for explaining why only those responses of agents that can be governed by reason are subject to normative reasons. This we saw in chapter 3 in our discussion of how reasons for motivation are reasons for action, as well as our discussion

in chapter 5 of how the scope of reasons differs from the scope of moral responsibility. We also saw something along these lines at the end of chapter 5, where we suggested that it appears to make no sense to mandate feelings from Sumatrans (or for them to mandate feelings form us) if they (or we) are genuinely psychologically incapable of having them.

We can deem it good or bad that people (or if you like "their bodies") do all sorts of things that are not under the scope of influence of their normative judgments. Thus if the mad scientist seizes control of my arm and uses it to strangle people, then the closing of my hand around someone's neck is something that we should hope doesn't happen and oppose if we can, but it is not something that I have reason to omit. In the same way, if I am genuinely psychologically incapable of feeling honor bound*, it seems to make no sense for a Sumatran who knows this to think that I have reason to feel honor bound* not to kill people. Surely what I should feel is obligated (or perhaps honor bound) not to kill people, for that is what our shared ethical reasoning can lead me to feel in the same way it can lead the Sumatran to feel honor bound* not to kill.

Moreover, that I don't have reason to do things with my hand that are completely out of my control, and that I don't have reason to feel things that I am psychologically incapable of feeling, appear to be conceptual rather than substantive normative truths. They do not look like the sorts of truths that we turn up in our deliberations, or by using our reflective equilibrium methods to capture such intuitions as those about who we should feel obligated not to kill and what sorts of things are worth pursuing and avoiding. Rather, we seem to get at these truths about the scope of reasons by making sense of our very concept of a normative reason for a response, which can be coherently applied to ever so many responses but not, it would seem, responses that I cannot control with my normative thoughts. But we can explain this if the notion of having reason to have a response is that of a response that one can correctly reason one's way into having. For doing things with my arm that I simply cannot do and having emotions that I'm not even set up to have are things that I cannot reason my way to doing or feeling on account of the fact that they are not within the scope of causal influence of my normative judgments.

Norm Descriptivism provides us with a straightforward explanation of why having reason to do something requires being able to do it as a result of judging that one

has such reasons. According to this view, having reason to do something is a matter of accepting fundamental norms that prescribe doing it. But part of what it is to accept such norms is for the states that constitute representations of them under a deliberative mode of presentation to directly influence motivation and action accordingly. Moreover, one cannot have representations of one's norms under this mode of presentation unless one accepts norms of which they are representations.[185] So according to Norm Descriptivism, the fact that beings like infants and sharks accept no norms entails both that they lack reasons to do things and that they cannot judge that they have them. Similarly too, the fact that I cannot have an alien emotion entails that I cannot accept norms that prescribe it, and the fact that I have lost use of my arm entails that my norms can no longer prescribe doing things with it.

But proponents of the independence of an agent's reasons from the fundamental norms she accepts seem unable to explain why an entity's having reasons to do things is dependent upon her ability to judge that she has them. As such, these theorists seem unable to explain why it is incoherent to attribute reasons to entities like infants, sharks, and volcanoes, or to think that I should do things that I cannot do or feel things that I cannot feel. If an entity's having reason to do something is a property that is analytically independent of its psychological states – like the property of maximizing happiness by so acting - why does it seem incoherent to attribute this property to entities that cannot represent and respond to it, like infants, sharks, and volcanoes? Why is it incoherent to attribute the property to responses that I cannot control through my normative thought, like letting go with my hand? One might, of course, just stipulate that the property of having reason to do something is the possession of a certain non-psychological property (like maximizing happiness by so acting) and being such that one can judge that one has it and respond in light of the judgment. But this accommodates the phenomenon without explaining it, and in fact makes normative facts dependent upon psychology in a much less principled way than Norm Descriptivism.

If one can accept norms for the circumstances of beings like infants and sharks who cannot make judgments about their reasons, Norm Expressivism and Relativism

[185] Because, as we have seen, these states have this content in virtue of bearing the right nomic relationship to (among other things) states of norm acceptance.

entail that one can coherently judge that they are subject to reasons. The Expressivist or relativist might try to prevent this by requiring, for instance, that to accept norms that prescribe having a certain response in a circumstance, one must take that circumstance to be such that one can accept norms in it, or norms that prescribe having or not having that response.[186] But this looks *ad hoc*. Why can one accept norms that prescribe having a response in a circumstance only if one could accept norms (for or against the response) in that circumstance? This does not seem to fall out of an account of shallow norm acceptance. It rather looks tacked-on by the Expressivist or Relativist to accommodate intuitions about the incoherence of certain judgments that her view cannot explain.

The expressivist or relativist may face similar problems when it comes to alien emotions and responses beyond our control. We have seen reasons to think that we cannot deeply accept norms that prescribe responses that we cannot have, and these might carry over to shallowly accepted norms too. This means that the expressivist and relativist can explain why it would be incoherent for me now to think that I should feel an alien emotion or that I should perform an action that is beyond my intentional control. But things may get funny when expressivists and relativists start talking about what other agents have reason to feel and do. For when the expressivist or relativist interprets thoughts about other agents' reasons as thoughts about how we should respond ourselves "in the agent's circumstances," what individuates the agent's circumstances? Some expressivists at least are against individuating them in such a way that they make allowance for what emotions other people are capable of feeling.[187] This flies in the face of intuitions to the effect that I can't have reasons to have alien emotions, and that Sumatrans (if their emotions are different) can't have reasons to have ours. But what might be worse, if we can't make allowance for what emotions people are capable of feeling, why should we make allowance for what people are capable of doing? If no mad scientist is controlling my right arm, and I can talk about my having reason not to omit choking someone with it, why can't I go on to say that other people have reason omit strangling people with their right arms, even if a mad scientist has control of them? Why

[186]I am grateful Allan Gibbard for very helpful discussion of these matters.
[187] Allan Gibbard, personal correspondence.

should the fact that the mad scientist is controlling the arms of others individuate their circumstances in a way that what emotions they are capable of feeling does not?

In fact, I suspect that the commitment of the expressivist and relativist to a strong version of the platitude about endorsement may favor their inviduating circumstances in such a way that they get the wrong results about other people's reasons not to strangle. For the idea we saw there was that if a consideration is a reason for one agent it must be a reason for another agent unless there is some strenghener or weakener that differentiates their situations that would make their reasons the same if it were present in both cases. But the fact that I cannot use my arm does not disable my reasons not to strange in the same way that the fact that a promise was given under duress disables my reasons to keep it. Facts about whether I can use my arm seem to be the sort of thing that determines what the justificatory framework applies to rather than what counts for what within that framework. But according to the alleged platitude about endorsement, it is the "what counts for what" considerations that determine sameness of reasons across sameness of circumstances. Since there is no difference between you who can use your arm and the person who cannot in terms of these kinds of reasons, it looks like the expressivist and relativist will have to say that, just as you can think about what to feel in Sumatran circumstances independently of what they can feel ("since what you can feel is no reason for or against feeling anything"), so too you can think about what to do in the circumstances of a man who has lost control of his arm to a mad scientist independently of what he can do (since whether you can use your arm is no reason for or against using it either).

I think, then, that Norm Descriptivism provides us with a better explanation than its opponents of the incoherence of holding entities like infants, sharks, and volcanoes subject to normative reasons. Similarly, I think that it provides us with a better explanation of why we can only have reason to do, think, or feel that which we can do, think, or feel in the sense of that action, thought, or feeling's being within the scope of influence of our normative thinking. This, I believe, is important evidence that Norm Descriptivism has the best explanation of the distinction between merely judging that we should oppose an entity's doing something and judging moreover that it has reason not to

do it. If, as I suspect, Norm Descriptivism has the best explanation of this distinction, then it is in a good position to debunk intuitions that it is coherent to judge that an agent has reason to do other than what her fundamental norms prescribe in terms of a lack of appreciation of what exactly her having reason to do something amounts to.

6.8. Norm Descriptivism and Ethics

I have thus argued in favor of understanding normative judgments, including judgments about which attitudes are fitting, in terms of the acceptance of norms for attitudes and (consequently) actions. Since these analyses do not rely upon the ethical concepts the fitting attitude analyst is seeking to analyze in giving an account of fittingness assessments, they provide us with a solution to the WKR problem. I have also argued that Norm Descriptivism is the best such analysis for the job. The Norm Descriptivist theory of fittingness, however, makes vivid a bit of unfinished business that actually applies to any fitting attitude analysis of ethical concepts. This is that, as Gibbard (1998, 253) notes, which attitudes it is fitting for an agent to have depends upon who the agent is and what her circumstances are like, whereas the instantiation of ethical and evaluative concepts like GOODNESS and MORAL BLAMEWORTHINESS do not. I shall conclude with a brief discussion of what the fitting attitude analyst can do to tie up this last loose end, particularly if she embraces Norm Descriptivism as theory of what it is for it to be fitting for an agent to have an attitude.

The problem of whose reasons for attitudes are spoken of in fitting attitude analyses of ethical and evaluative concepts is faced by any fitting attitude analyst, Norm Descriptivist or not, for at least *some* ethical and evaluative concepts. Perhaps the most obvious problem concept for everyone is that of a GOOD STATE OF AFFAIRS. Even if we are Expressivists or Platonists about what it is to judge an attitude fitting, we should concede that it is at least coherent to think that what states one should desire depends upon what one's circumstances are like. If, for instance, S_1 is a state of affairs where Jack will die but Bill will live, and S_2 is a state of affairs where Bill will die and Jack will live, it seems at least coherent to judge that it is fitting for Jack to prefer S_2 to S_1 while it is

fitting for Bill to prefer S_1 to S_2. Similarly, as Gibbard (1998) points out, it seems coherent to think that it is fitting for fans of rival football teams to desire states in which it is their team that wins, and fitting for rival job applicants each to desire a state of affairs in which she is the one hired for a job.

If, then, it is fitting for various agents to desire different states of affairs, whose reasons for desire does the fitting attitude analyst think we are talking about when we judge states of affairs to be good? Gibbard's (1998) solution is to view the content of the concept we express by 'good state of affairs' as context sensitive:

> How does [our concept of a GOOD or BETTER EVENT] work? There is a kind of impartiality built into the concept: Not that it requires an impartiality that is universal: fans of the Michigan Wolverines football team can exclaim together about what's good and what's bad as a game progresses, without bringing in the standpoint of anyone else. But the talk of good and bad is still neutral among participants in the conversation. The same talk would be tendentious in dealings with people from Ohio. It couldn't be regarded as seriously defensible, or if it were, the grounds would have to be different.
>
> Likewise I can think to myself about good and bad developments as, say, I compete for a job – without having to decide whether my getting the job is best on an impartial view of things. My conversational group has then shrunk to one – though with luck I may be able to rope in a few friends to share my good news and bad. They can't, though, be equally friends of another competitor, or I'll have to qualify my language or make the conversation awkward for them. (Gibbard 1998, 254-255).

The idea, then, is that to say that a state of affairs is good is to say that it is fittingly desired by everyone common to the conversational group. One way to spell out Gibbard's contextualist suggestion here would be to understand the concept of a GOOD STATE OF AFFAIRS to be equivalent to that of A STATE OF AFFAIRS THAT IT IS FITTING FOR EVERYONE TO DESIRE, but to understand the universal quantifier in the *analysans* to have a contextually determined domain restriction[188] that determines that (at least) both the speaker and audience are in the domain, rather than a literally unrestricted domain (or a domain ranging over all conceivable agents).

This kind of solution can also be applied to explain how the modal behavior of judgments about goodness can diverge from the modal behavior of judgments about what it is fitting for a particular person to desire. Thus, suppose that two young children with big egos, Billy and Jimmy, are both competing in a chess tournament, with their parents,

[188] Of the kind discussed by Stanley and Williamson (1995) and Stanley and Szabo (2000).

Bugsy, Suzy, Fred, and Judy watching. Both sets of parents know that whichever child loses will throw a tantrum and life will be hell for that set of parents for the next three days. As the game unfolds, it seems that Bugsy might say to Suzy 'It sure will be good if Billy remembers not to bring his queen out too early – though of course, if I were in Fred and Judy's shoes, I'd have every reason to hope that he forgets!' But it would sound completely bizarre for Bugsy to say 'if I were in Fred and Judy's shoes, it'd be bad for Billy to remember not to bring out his queen early'. It seems that it we could easily have had reason to prefer different states of affairs if our circumstances had been otherwise, but that our circumstances being otherwise would not change what is good. This the fitting attitude can explain by including an actuality operator in her analysis of judgments about good states of affairs: to judge a state good is to judge that the contextually salient set of agents *actually* have reason to desire it.[189]

Now, if Norm Descriptivism is correct, which attitudes it is fitting for an agent to have always depends upon what is prescribed by the most fundamental norms she deeply accepts. Thus, for the Norm Descriptivist, the dependence of what attitudes it is fitting to have on which agent we are talking about goes not only for what it is fitting to desire, but indeed for what it is fitting to feel angry at, fitting to feel obligated to do, and so on. Perhaps, as I think is the case, we psychologically typical adult humans all accept systems of norms that prescribe anger and scorn towards all and only the same actions,

[189] There is, however, a complication with how this sort of analysis interacts with how we explain agents' behavior in terms of their judgments about what is good. For we might want to be able to say things like: 'If Bugsy had thought it good for Jimmy to win, then he would have tried to distract Billy', where this clearly does not mean 'If Bugsy had thought that Bugsy as he actually is had reason to desire Jimmy to win, then he would have tried to distract Billy'. Yet if we thought about the contextualist proposal on the model of indexicals, this is what we would get, since using an indexical in an embedded context makes it refer to what it refers to in the mouth of the speaker, not in the mouth of the person who bears the relevant attitude towards the proposition indicated by the 'that'-clause. Thus, if I say 'Bill went to the store because he thought that I was out of bread' , I end up saying that Bill went to the store because he thought that I, the speaker, was out of bread, not because he thought that he, the person who went to the store was out of bread.

I think, however, that the kind of contextual expression we can get with quantifier domain restrictions can work differently. Thus, when we say 'Bill thinks that every book is on the table', we can make salient either the books around here or the books around Bill. What we might be able to get with judgments about goodness is the salience of a set of agents that includes the judge as she is in the worlds in which she is judging when we attribute a goodness judgment, but that includes the salience of us and the other agents as they actually are in our world when we make a goodness judgment. But whether we have any principled reason to believe in this kind of pattern of salience is something that I think needs further investigation. I am grateful to Josh Dever for alerting me to these difficulties.

indeed regardless of our circumstances.[190] But we can of course imagine agents on a distant planet who fundamentally accepted norms that prescribed that they have these attitudes towards any bizarre patterns of actions we care to stipulate. Moreover, there may (though there may not) actually be certain individuals, who we might call "severe sociopaths," who not only have difficulty feeling emotions like guilt, shame, feeling obligated, feeling honor bound, and perhaps also the outrage that is guilt's third person counterpart, but are in fact psychologically incapable of having these attitudes. Deeply accepting norms for a response requires that judgments about when these norms prescribe the response exert causal pressure in the direction of one's having it. But causal pressure cannot be exerted in the direction of having an attitude that one is psychologically incapable of having. Thus, if there are such individuals as severe sociopaths for whom feelings of obligation, guilt, and outrage are genuine psychic impossibilities, the fundamental norms they deeply accept cannot prescribe these attitudes, however much these fundamental norms may be otherwise like ours (*qua* prescriptions for prudential desires and beliefs, for instance). Norm Descriptivism would thus entail that it is not fitting for these agents to have such attitudes.[191]

So for such analyses as that of moral blameworthiness into what it is fitting to feel anger at others for doing (and guilt for doing oneself), the Norm Descriptivist does, where the Norm Expressivist or Relativist may not, face the question of whose reasons for anger we are talking about. But all the Norm Descriptivist needs to do is to apply the contextualist strategy that Gibbard (1998) pioneered for fitting attitude analyses of good states of affairs to all other fitting attitude analyses. For instance, on this contexutalist account, the concept of MORAL BLAMEWORTHINESS is that of WHAT IT IS FITTING FOR THE ACTOR TO FEEL GUILT FOR DOING, AND WHAT IT IS FITTING FOR EVERYONE ONE TO FEEL

[190] Which, as you might have gleaned from Chapter 4, I actually suspect to be such that they prescribe that we never have these attitudes in our actual circumstances.

[191] I should stress, however, that that this conclusion holds only for those agents I am here calling severe sociopaths who are psychologically incapable of emotions in question. The term 'sociopath' is often used in a much more general way, to refer to someone who may exhibit enough of the canonical anti-social signs and symptoms on a "sociopath checklist," one of which is typically shallowness of affect or lack of guilt or remorse, but not necessarily psychological incapability of these emotions. Mere lack of feeling certain emotions is of course no guarantee that one is unable to feel them or that one does not deeply accept causally influential principles that prescribe feeling them (the causal influence of which may be overwhelmed by other causal forces, or is exercised only in the direction of suppressing the emotions due to false appearances that they are not in one's circumstances warranted).

ANGER AT HIM FOR DOING, but where the universal quantifier EVERYONE has an implicit domain restriction to at least the person tokening the concept and her intended audience. Since this audience can (and typically will) exclude aliens and known severe sociopaths, the existence of these beings and their lack of the relevant reasons poses no threat to our getting on with the business of making assessments of moral blameworthiness and ascribing the reasons for anger to a set of agents that excludes these beings.

Now on the understanding of moral wrongness and moral reasons we have developed, Norm Descriptivism entails that agents who fundamentally accept no norms for moral emotions like feelings of obligation lack these reasons. For whether it is morally wrong to do something requires that there be a rational mandate for one to feel obligated not to do it oneself (quite independently of what the people assessing one's conduct are rationally mandated or permitted to feel). For reasons we have seen, if there are sociopaths or there were space aliens who are psychologically incapable of feeling obligated, it will be impossible for them to deeply accept fundamental norms that have a propensity to cause them to have these attitudes. As such, these beings would lack moral reasons, and their actions could never be morally wrong. Nor would it simply be that the severe sociopaths and aliens would be like the Sumatrans in that we can say that their actions have an ethical status that has the same practical purport as wrongness, like "being ethically out." For severe sociopaths appear to be incapable of any kind of aversive valenced attitude that that would tend to deter them from doing the same kind of things that we deem wrong or lowly for anything like the same reasons. Absent fear or aversions to actually being sanctioned, they would seem to be capable of no deterrent attitude to doing horrible things that we might otherwise hope to be prescribed by their deeply accepted fundamental norms. Similar remarks could go for any aliens we might care to imagine up.

It is of course consistent with this that any horrible things these agents do are still bad and such that we (non-sociopaths) have reason to prevent them from doing them. We are quite used to the idea that horrible things done by many beings are bad and to be prevented without their being in any way wrongful. This is surely our attitude towards such things as sharks attacking our friends and coyotes attacking our companion animals.

Certain space aliens and severe sociopaths would be different from sharks and coyotes in that they are agents who are subject to some reasons, like reasons for belief and prudent action. But if the foregoing is correct, the best explanation of why sharks and coyotes cannot wrongfully harm is that they cannot reason their way to refraining from inflicting harm out of judgments that they should feel obligated not to do so. This explanation equally entails that sociopaths cannot wrongfully harm if they cannot reason their way to moral emotions. The fact that they can reason their way to other attitudes is irrelevant.

Suppose, however, that there was in a distant galaxy a space alien who accepted norms that mandated that he feel obligated to torture innocent humans whenever the chance arises, and that should he fail to do this he should feel guilt for the failure. It would be one thing for us to say that the alien has reason to torture innocent humans. It would also be one thing for us to say of her, as we might of sharks, coyotes, and sociopaths, that she does no moral wrong in torturing innocent humans. But it seems utterly incorrect to say that her failure to torture innocent humans would be morally wrong.

Of course, the Norm Descriptivist is not in any danger of having to say that the space alien's failure to torture innocent humans is *blameworthy*, even if she is fully responsible for her actions. For as we have seen, the Norm Descriptivist will say that to judge the alien's conduct blameworthy is to judge that all agents in the relevant domain of quantification would be justified in feeling angry at the space alien for failing to torture innocent humans whenever the chance arises. But were we to make this judgment, the domain of quantification would include us ourselves. And it is pretty obvious (as apparently non-debunked intuitions testify) that we do not accept fundamental norms that prescribe feeling angry at the space alien for failing to do this. Since we would not be justified in feeling angry at the space alien for failing to torture innocent humans just for fun, so we would speak falsely were we to judge her blameworthy and say something that entailed that we did. The space alien may have reason to feel guilt for her failure, but since we do not have reason to feel anger at her for it, the failure is not something that we could correctly call blameworthy.

So much, perhaps, for blameworthiness, but what of wrongness? Our fitting attitude analysis of moral blameworthiness wears on its face an open place for the Norm

Descriptivist to tell a contextualist story, in the form of the "other agents" on the part of whom fitting anger is held to constitute an act's blameworthiness. Our fitting attitude analysis of moral wrongness, on the other hand, seems to speak only of the reasons for attitudes of the agent whose conduct is being judged to be wrong. If all the analysis says it takes for an agent's doing something to be morally wrong is for her to have reason to feel obligated not to perform the act, how can the Norm Descriptivist avoid calling wrongful a failure to torture innocent humans on the part of an agent whose fundamental norms *ex hypothesi* prescribe that she feel obligated never so to fail?

The answer, I think, comes in the form of a stronger conceptual connection between moral wrongness and moral blameworthiness than that we have explicitly suggested. As we have seen in spades, not everything that is morally wrong is morally blameworthy – there are indeed exculpated wrongs. A person might do something she should feel obligated not to do, but if she was less than fully responsible for her action it might be unfair for others to feel angry at her for doing it. In Chapter 5 we considered understanding the foregoing notion of responsibility in performing an action as itself a normative notion – namely that of being such that should your action be the kind of thing you should feel obligated not to perform before the fact, others will be justified in being angry at you (and you should feel guilt) for doing it after the fact. I believe that there are many advantages to this approach, but they are not advantages of which it would be wise for a Norm Descriptivist to avail herself. The Norm Descriptivist would do better to treat the question of moral responsibility as something else – perhaps it just is the question of control, and perhaps what I called 'Barbarism' is actually an incoherent position.[192]

---

[192] Though perhaps it is not; the question of what control amounts to may be an open conceptual question but a conceptual question nonetheless. We may not yet know who has the correct analysis, but we know that whosever loses will be losing a battle about what is coherent, not just about what is substantively normatively some way (though of course the two are related, for we have some choice which concepts to bother think in terms of).

There may also be a way in which the norm descriptivist can attempt to construe the notion of responsibility as a normative notion in a way more fine-grained than simply the difference between what the actor's must feel obligated not to do before the fact and what others may be angry at her for doing afterwards. The Norm Descriptivist might construe the question of responsibility as something like the question of whether, assuming that the actor accepted the same fundamental norms as us, our norms prescribe feeling angry at her after the fact. But even if the Norm Descriptivist can in this sort of way allow the notion of excuse to be a normative notion, it does not, I think, alter her basic theoretical situation. In order to capture intuitions about the falsity of talk of the wrongness of the conduct of the alien described in the text, she will need to explicitly inset the notion of BLAMEWORTHY ABSENT EXCUSE into the notion of moral wrongness, whether EXCUSE is a normative notion or not.

What the Norm Descriptivist should say, then, is that to judge an act wrong is more than just to judge that its author should feel obligated not to perform it. She should also say that judging an act morally wrong involves judging that it would be blameworthy if the actor performed it and was fully responsible. She should thus embrace the following version of our fitting attitude analysis of moral wrongness:

> **Norm Descriptivist's Fitting Attitude Analysis of Moral Wrongness**:
> To judge that agent *A*'s act of ϕ-ing is morally wrong is to judge that (1) unless *A* is already going to refrain from ϕ-ing anyway, it is rationally mandatory for *A* to feel obligated not to ϕ, and (2) If *A* does responsibly ϕ, then *A*'s ϕ-ing is morally blameworthy – i.e. it is fitting for *A* to feel guilty for having ϕed, and fitting for others to be outraged at or resentful of *A* for having ϕed.

Equipped with this, the Norm Descriptivist is able to explain how not only judgments of moral blameworthiness but also judgments of moral wrongness are beholden to more than the prescriptions of the norms of the agent whose moral status is being assessed. Due to the kind of context sensitivity exhibited by judgments of blameworthiness and their conceptual relation to those of moral wrongness, such assessments require for their truth a confluence between (at least) the prescriptions of the fundamental norms of the agent assessed and those of the agent doing the assessing.

As we have seen, our judgments to the effect that the space alien is morally blameworthy for failing to torture innocent humans whenever she can is false because we lack reason to be angry with the alien for failing to do this and we are in the domain of quantification of agents, our judgments of whose reasons for anger in part constitute our judgment of blameworthiness. We lack such reasons to be angry with the space alien for failing to torture innocent humans whenever she can even if the space alien is fully responsible for her failure to do so. Hence, any judgment on our part to the effect that IF THE SPACE ALIEN RESPONSIBLY FAILS TO TORTURE INNOCENT HUMANS JUST FOR FUN, THEN HER FAILURE IS BLAMEWORTHY would have to be false. But since, by the Norm Descriptivist's fitting attitude analysis of moral wrongness, a judgment on our part that it would be morally wrong for the space alien to fail to torture innocent humans just for fun

entails this judgment about blameworthiness, our judgment that her failure would be wrong must similarly be false.

One thing that Norm Descriptivism and the universal application of the above kind of contextualism to fitting attitude analyses would entail is that ethical and evaluative talk cannot be used truly in conversations with agents whose deeply accepted fundamental norms prescribe sufficiently different attitudes than one's own. This, I think, is as it should be. As Gibbard's examples of the Michigan and Ohio fans illustrates, evaluative talk intuitively breaks down where shared reasons for attitudes break down. The evidence in favor of Norm Descriptivism counts in favor of the view that the sharing of reasons for attitudes breaks down just where the sharing of deeply accepted norms that prescribe them breaks down. Were such a breakdown occur, we would have no choice but to fall back on thought about what is fitting for whom to have which attitudes towards, without any presupposition that the same things warrant the same attitudes on the part of all agents. But of course, this retreat from ethical thought to thought of which attitudes are fitting for whom does not really give up anything central to our normative thought if, as the contexutalist fitting attitude analyst has it, the former is simply the latter with a presupposition of shared reasons for feeling.

Appendix:

Proof That There is Conclusive Reason Not to Do What is Morally Wrong

The Stock of Conceptual Truths

Fitting Attitude Analysis of Moral Wrongness*:
To judge that agent *A*'s act of ϕ-ing is morally wrong is to judge that, unless *A* is already going to refrain from ϕ-ing anyway, it is rationally mandatory for *A* to feel obligated not to ϕ

Most-Motivation-Action Principle (3):
(3) If it is rationally mandatory for one to be most strongly motivated to do *X*, then one has conclusive reason to do *X* (i.e. it is rationally impermissible for one to fail to do *X*).

Contour Thesis:
If it is rationally mandatory for agent *A* to feel obligated to do *X*, then it is rationally mandatory for *A* to be most strongly motivated to do *X*.

Motivation Partition Principle:
If agent *A* is such that every motivational state that it would be rationally permissible for *A* to be in involves being most strongly motivated to do *Y*, then it is rationally mandatory for *A* to be most strongly motivated to do *Y*

The Proof

For any agent *A* and action *X*,

(1) If it would be morally wrong for *A* to do *X*, then either (C1) it would be morally wrong for *A* to do *X* and *A* is *not* already sufficiently deterred from doing *X*, or (C2) it would be morally wrong for *A* to do *X* and *A is* already sufficiently deterred from doing *X*

(2) If (C1), then it is rationally mandatory for *A* to feel obligated not to do *X* [Fitting Attitude Analysis of Wrongness*]

(3) If it is rationally mandatory for *A* to feel obligated not to do *X*, then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [Contour Thesis]

∴(4) If (C1), then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [2, 3]

(5) If (C2), then either (C21) *A*'s state of sufficient deterrence to doing *X* is justified, or (C22) *A*'s state of sufficient deterrence to doing *X* is not justified

(6) If (C21), then either (C211) *A*'s state of sufficient deterrence to doing *X* is rationally mandatory, or (C212) *A*'s state of sufficient deterrence to doing *X* is rationally optional

(7) If (C211), then, *ex hypothesi*, it is rationally mandatory for *A* to be most strongly motivated not to do *X*.

(8) If (C212), then there is some motivational state *M* such that *M* is a rationally permissible alternative to *A*'s actual state of motivation *M**, and for every such alternative state of motivation *M*, either (C2121) *M* is not a state of sufficient deterrence to doing *X* in the absence of *A*'s having to feel obligated not to do *X*, or (C2122) *M* is a

state of sufficient deterrence to doing *X* in the absence of *A*'s having to feel obligated not to do *X*,

(9) If doing *X* is morally wrong, then for every rationally permissible state *M* such that (C2121), *M* involves feeling obligated not to do *X* and *M* is a state of being most strongly motivated not to do *X* [Fitting Attitude Analysis of Wrongness*, Contour Thesis]

(10) For every motivational state *M* such that (C2122), *M* involves, *ex hypothesi*, being most strongly motivated not to do *X*

∴ (11) If doing *X* is morally wrong and (C212), then for every rationally optional alternative *M* to *A*'s actual state of motivation *M**, *M* involves being most strongly motivated not to do *X* [8-10]

(12) If doing *X* is morally wrong and (C212), then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [11, Motivation Partition Principle]

∴ (13) If doing *X* is morally wrong and (C21), then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [7, 12]

(14) If (C22), then it is rationally mandatory for *A* to occupy a different state of overall motivation, and for every rationally permissible state *M* such that *M* is a rationally permissible alternative to *A*'s actual state of motivation *M**, either (C221) *M* is not a state of sufficient deterrence to doing *X* in the absence of *A*'s having to feel obligated not to do *X*, or (C222) *M* is a state of sufficient deterrence to doing *X* in the absence of *A*'s having to feel obligated not to do *X*

(15) If doing *X* is morally wrong, then for every rationally permissible state *M* such that (C221), *M* involves feeling obligated not to do *X* and *M* is a state of being most strongly motivated not to do *X* [Fitting Attitude Analysis of Wrongness*, Contour Thesis]

(16) For every motivational state *M* such that (C222), *M* involves, *ex hypothesi*, being most strongly motivated not to do *X*.

∴ (17) If doing *X* is morally wrong and (C22), then for every rationally optional alternative *M* to *A*'s actual state of motivation *M**, *M* involves being most strongly motivated not to do *X* [14-16]

(18) If doing *X* is morally wrong and (C22), then it is rationally mandatory for *X* to be most strongly motivated not to do *X* [17, Motivation Partition Principle]

∴ (19) If (C2), then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [13, 18]

∴ (20) If it would be morally wrong for *A* to do *X*, then it is rationally mandatory for *A* to be most strongly motivated not to do *X* [4, 19]

(21) If it is rationally mandatory for *A* to be most strongly motivated not to do *X*, then *A* has conclusive reason not to do *X* [Most-Motivation-Action Principle]

∴ (22) If it would be morally wrong for *A* to do *X*, which is to say *A* is morally obligated not to do *X*, then *A* has conclusive reason not to do *X* [20, 21]

Bibliography

Adams, Fred and Kenneth Aizawa. 1992. "X" Means X: Semantics Fodor-Style. *Minds and Machines*, 2: 175-183.

Aristotle. ca. 350 B.C.E. *Nicomachean Ethics*. Ross, W.D. (trans) revised by J.O. Urmson, in *The Complete Works of Aristotle*, The Revised Oxford Translation, vol. 2, Jonathan Barnes, ed., Princeton: Princeton University Press, 1984.

Axelrod, Robert and William D. Hamilton. 1981. The Evolution of Cooperation. Science, 211:1390-1396.

Ayer, Alfred J. 1936. *Language, Truth, and Logic.* 2nd ed. London: Victor Gollancz.

Blackburn, Simon. 1988. "How to Be an Ethical Antirealist," *Midwest Studies in Philosophy* 12: 361-375. Reprinted in n S. Darwall, A. Gibbard, and P. Railton, eds., *Moral Discourse and Practice*. New York: Oxford University Press.

Blackburn, Simon.1998. *Ruling Passions*. Oxford: Clarendon Press.

Boghossian, Paul and Christopher Peacocke (eds.). 2000. Introduction. *New Essays on the A Priori*. Oxford: Clarendon Press.

BonJour, Laurence. 1998. *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

Boorse, Christopher and Roy A. Sorensen. 1988. Ducking Harm. Journal of Philosophy, 85: 115-134.

Boyd, Richard, 1988. How to Be a Moral Realist. In *Essays on Moral Realism*, edited by G. Sayre-McCord, 181-228. Reprinted in n S. Darwall, A. Gibbard, and P. Railton, eds., *Moral Discourse and Practice*. New York: Oxford University Press.

Boyd, R. and P.J. Richardson. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13: 171-195.

Braddon-Mitchell, David. 2005. The Subsumption of Reference. *British Journal for the Philosophy of Science*, 56:157-178.

Braddon-Mitchell, David and Frank Jackson. 2007. *Philosophy of Mind and Cognition: An Introduction*. Malden, MA: Blackwell.

Brandt, Richard B. 1946. Moral Valuation. *Ethics*, 56: 106-121.

Brandt, Richard B. 1959. *Ethical Theory*. Englewood Cliffs, N.J: Prentice Hall.

Brandt, Richard B. 1979. *A Theory of the Good and the Right*. Amherst, New York: Prometheus Books.

Bratman, Michael E. 1987. *Intention, Plans, and Practical Reason*. Cambridge, Massachusetts: Harvard University Press.

Braun, Jochen. 2000. *Nature* 408: 154-155.

Carritt, E. F. 1947. *Ethical and Political Thinking*. Oxford: Oxford University Press.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford Univeristy Press.

Chalmers, David J. and Frank Jackson. 2001. Conceptual Analysis and Reductive Explanation. *The Philosophical Review*, 110: 315-360.

Copp, David. 1995. *Morality, Normativity, and Society*. New York: Oxford University Press.

Dancy, Jonathan. 2004. *Ethics without Principles*. Oxford: Clarendon Press.

Daniels, Norman. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy*, 76:5: 256-82.

Daniels, Norman. 1980. On Some Methods of Ethics and Linguistics. *Philosophical Studies*, 37: 21-36.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam's Sons.

D'Arms, Justin and Daniel Jacobson. 1994. Expressivism, Morality, and the Emotions. *Ethics*, 104: 739-765.

D'Arms, Justin and Daniel Jacobson. 2003. The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism). *Philosophy: The Journal of the Royal Institute of Philosophy*, 52 (suppl.): 127-145.

D'Arms, Justin and Daniel Jacobson. 2000. The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research*: 65-90.

Darwall, Stephen L. 1983. *Impartial Reason.* Ithaca: Cornell University Press.

Darwall, Stephen L. 2002. *Welfare and Rational Care.* Princeton: Princeton University Press.
Darwall, Stephen L. 2003. Moore, Normativity, and Intrinsic Value. *Ethics*,113: 468-489.

Darwall, Stephen L. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge: Harvard University Press.

Davidson, Donald 1982. Rational animals. *Dialectica,* 36: 317-327.

de Waal, Frans B.M. 1991. The Chimpanzee's Sense of Social Regularity and its Relation to the Human Sense of Justice. *American Behavioral Scientist,* 34 (3): 335-349.

de Waal, Frans B.M. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge: Harvard University Press.

Dretske, F.I. 1981. *Knowledge and the Flow of Information*. Cambridge: MIT Press.

Dretske, F.I. 1988. *Explaining Behavior*, Cambridge, MA: Bradford, MIT.

Enoch, David. 2007. An Outline of an Argument for Robust Metanormative Realism. In Russ Shafer-Landau, ed., *Oxford Studies in Metaethics: Volume 2*. Oxford: Oxford University Press.

Egan, Andy. 2006. Quasi-Realism and Fundamental Moral Error. *Australasian Journal of Philosophy*, 85:2: 205-219.

Ewing, A.C. 1939. A Suggested Non-Naturalsitic Analysis of Good. *Mind,* 48: 1-22.

Falk, W.D. 1948. 'Ought' and Motivation. *Proceedings of the Aristotelian Society*, 48: 111-138. Reprinted in *Ought, Reasons, and Morality: the Collected papers of W.D. Falk,* by W.D. Falk. Ithica: Cornell University Press.

Falk, W.D. 1963. "Action-Guiding Reasons." *The Journal of Philosophy,* 60: 703-18.

Falk, W.D. 1986. "On Learning about Reasons." In Falk, *Ought Reasons, and Morality: The Collected Papers of W.D. Falk.* Ithica: Cornell University Press.

Fessler, Daniel M. T. and Kevin Haley. 2003. The Strategy of Affect: Emotions in Human Cooperation. In P. Hammerstein (ed.) *Genetic and Cultural Evolution of Cooperation*, Cambridge, MA: MIT Press.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

Findlay, Stephen. 2007. Responding to Normativity. In Russ Shafer-Landau, ed., *Oxford Studies in Metaethics: Volume 2*. Oxford: Oxford University Press.

Firth, Roderick. 1952. Ethical Absolutism and the Ideal Observer Theory. *Philosophy and Phenomenological Research*, 12: 317-345.

Fischer, John Martin and Mark Ravizza. 1992. *Ethics: Problems and Principles*. Orlando: Harcourt Brace Jovanovich College Publishers.

Fodor, Jerry A. 1987. *Psychosemantics*: the Problem of Meaning in the Philosophy of Mind. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays*. Cambridge, Massachusetts: MIT Press.

Foot, Philippa. 1959. Moral Beliefs. *Proceedings of the Aristotelian Society*, 59:83-104.

Foot, Philippa. 1963. Hume on Moral Judgment. David Pears, ed. *David Hume*, London. Reprinted in P. Foot, *Virtues and Vices*, Oxford: Blackwell, 1978.

Foot, Philippa. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, 5: 5-15.

Frankena, W.K. 1939. The Naturalistic Fallacy. *Mind*, 48: 1939.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings.* Cambridge, Massachusetts: Harvard University Press.

Gibbard, Allan. 1998. Preference and Preferability. In Christoph Fehige and Ulla Wessels, eds., *Preferences*. Berlin: Walter de Gruyter.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, Massachusetts: Harvard University Press.

Gibbard, Allan. 2006. Reply to Critics. *Philosophy and Phenomenological Research*. 72: 729 -744.

Gibbard, Allan. 2007. Thinking How to Live Together. In Grethe B. Peterson, ed., *The Tanner Lectures on Human Values*, vol. **27** (Salt Lake City: University of Utah Press), 165–226.

Gintis, H. S. Bowles, R. Boyd, and E. Fehr. 2003. Explaining Altruistic Behavior in Humans. Evolution and Human Behavior, 24: 153-172.

Goodman, Nelson. 1954. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Gould, S.J. and R.C. Lewontin. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society: Biological Sciences*, 205: 581-98.

Greenspan, Patricia. 1988. *Emotions and Reason: An Inquiry into Emotional Justification*. London: Routledge& Kegan Paul.

Hall, R. J. 2008. If it itches, scratch! *Australasian Journal of Philosophy*, 86: 525–535.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

Hare, R. M., 1952, *The Language of Morals*, Oxford: Clarendon Press.

Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Free Press.

Hobbes, Thomas. 1651. *Leviathan*. Edited by Edwin Curley. Indianapolis: Hackett Publishing Company, 1994.

Hume, David. 1739. A Treatise of Human Nature. L.A. Selby Bigge (ed), Oxford: Clarendon, 1978.

Jackson, F. 1982. Epiphenomenal Qualia. *Philosophical Quarterly*, 32: 127-136.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

James, William. 1890. *The Principles of Psychology*. 2 vols. New York: Dover.

James, William. 1897. *The Will to Believe and Other Essays in Popular Philosophy*. Cambridge, MA and London: Harvard University Press, 1979.

Joyce, R. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.

Kagan, Shelly. 1998. *Normative Ethics*. Boulder: Westview.

Kamm, Francis M. 1993. *Morality and Mortality.* Vol. 1. Oxford: Oxford University Press.

Kaplan, David. 1989. Demonstratives. In J. Almog, J. Perry, and H. Wettstein, eds., *Themes from Kaplan*. New York: Oxford University Press.

Kavka, Gregory S. 1983. The Toxin Puzzle. *Analysis*, 43: 33-36.

Kenny, Anthony. 1963. *Action, Emotion and Will*. London; New York: Routledge and Kegan Paul; Humanities Press.

Kitcher, Philip S. 1993. The Evolution of Human Altruism. *Journal of Philosophy*, 90: 497-516.

Kitcher, Philip S. 1998. Psychological Altruism, Evolutionary Origins, and Moral Rules. *Philosophical Studies*, 89: 283-316.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. New York: Cambridge University Press.

Krebs, Dennis. 2005. The Evolution of Morality. In D.M. Buss (ed) *The Handbook of Evolutionary Psychology*, Hoboken, NJ: John Wiley & Sons.

LeDoux, Joseph. 1998. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Lewis, David. 1970. How to Define Theoretical Terms. *Journal of Philosophy*, 67: 427-446.

Lewis, David. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, 50: 249-58.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

Lewis, David. 1979. Attitudes De Dicto and De Se. *Philosophical Review*, 88: 513-543.

Lewis, David. 1980. Mad Pain and Martian Pain. In N. Block (ed) *Readings in the Philosophy of Psychology*. Cambridge, MA: Harvard University Press.

Lewis, David. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, 63 (suppl.): 113-37.

Lewis, David. 1996. Desire as Belief II. *Mind*, 105: 303-313.

Lycan, W. G. 1987. *Consciousness*. Cambridge, Massachusetts: MIT Press.

Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong*, New York: Penguin.

Malm, Heidi. 1989. Killing, Letting Die, and Simple Conflicts. *Philosophy and Public Affairs*, 18: 238-258

Marr, D. 1980. *Vision*. Boston: Houghton Mifflin.

Mason, Jim and Peter Singer. 1990. *Animal Factories*. New York: Harmony Books.

McMahan, Jeff. 2000. Moral Intuition. In Hugh LaFollette, ed., *Blackwell Guide to Ethical Theory*. Oxford: Blackwell.

McMahan, Jeff. 2002. *The Ethics of Killing: Problems at the Margins of Life*. New York and Oxford: Oxford University Press.

McMahan, Jeff. 2003. Animals. In R.G. Frey and Christopher Wellman, eds., *Companion to Applied Ethics*. Oxford: Blackwell.

Mealey, Linda. 1995. The sociobiology of sociopathy: an integrated evolutionary model. *Behavioral and Brain Sciences* 18:523–99. Reprinted in Baron-Cohen, Simon, ed., *The Maladapted Mind*.

Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son, and Bourn.

Moore, G.E. 1903. *Principia Ethica*. New York: Cambridge University Press.

Moore, Michael S. 2000. The Moral Worth of Retribution. In Feinberg (ed) *Philosophy of Law*, Belmont, CA: Wadsworth.

Murphy, J. G., and J. Hampton. 1988. Forgiveness and Mercy. Cambridge: Cambridge University Press.

Nakhnikian, G. 1963. On the Naturalistic Fallacy. In Hector-Neri Castaneda and George Nakhnikian (eds), *Morality and the Language of Conduct*. Detroit: Wayne State University Press, 145–158.

Nichols, Shaun. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgments*. Oxford: Oxford University Pres.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. United States of America: Basic Books.

Nunner-Winkler, Gertrude and Beate Sodian. 1988. Children's Understanding of Moral Emotions. *Child Development*, 59: 1323-1338.

Olson, Jonas. 2004. Buck-Passing and the Wrong Kind of Reasons. *The Philosophical Quarterly* 54: 295-300.

Parfit, Derek. 1984. *Reasons and Persons*, Oxford: Oxford University Press.

Parfit, Derek. 2001. Rationality and Reasons. In *Exploring Practical Philosophy: From Action to Values*, edited by Dan Egonsson, Bjorn Petersson, Jonas Josefsson, and Toni Ronnow-Rasmussen, 17-41. Aldershot: Ashgate.

Perry, John. 1979. The Problem of the Essential Indexical. *Noûs*, 13: 3-20.

Plato. ca. 380 B.C.E. *Republic*. G.M.A. Grube (trans.), *Plato. The Republic*. revised by C.D.C. Reeve. Indianapolis: Hackett, 1992.

Pogge, Thomas W. 1995. Three Problems with Contractarian-Consequentialist Ways of Assessing Social Institutions. *Social Philosophy and Policy*, 12: 241-266.

Prichard, H.A. 1912. Does Moral Philosophy Rest on a Mistake? *Mind*, 21: 21-37

Prinz, Jesse. 2006. The Emotional Basis of Moral Judgments. *Philosophical Explorations*, 9: 29 -43.

Price, M.E., L. Cosmides, and J. Tooby. 2002. Punitive Sentiment as an Anti-Free Rider Psychological Device. Evolution and Human Behavior, 23: 203-231.

Putnam, Hilary. 1975. "The Meaning of 'Meaning.'" In Putnam, H. *Mind, Language and Reality*, Cambridge, Cambridge University Press, p.215-271. (First published in K. Gunderson, ed., *Language, Mind and Knowledge*, Minneapolis: University of Minnesota Press).

Quartz, Steven R. and Terrence J. Sejnowski. 2002. *Liars, Lovers, and Heroes: What the New Brain Science Reveals About How We Become Who We Are*. New York: Harper Collins Publishers.

Quine, W.V.O. 1951. Two Dogmas of Empiricism. *Philosophical Review*, 60: 20-43. Reprinted in Quine 1953, *From a Logical Point of View*, Cambridge, MA: Harvard University Press

Quine. W.V.O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quinn, Warren. 1993. Putting Rationality in its Place. In R. Frey and C. Morris, *Value, Welfare, and Morality*. New York: Cambridge University Press, 26-50. Reprinted in W. Quinn, *Morality and Action*, New York: Cambridge: Cambridge University Press, 228-255.

Rabinowicz, Wlodek and Toni Ronnow-Rasmussen. 2004. The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics* 114: 391-423.

Rabinowicz, Wlodek and Toni Ronnow-Rasmussen. 2006. Buck-Passing and the Right Kind of Reasons. *The Philosophical Quarterly* 56: 14-120.

Railton, Peter, 1986. Moral Realism. *Philosophical Review*, 95: 163-207.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, John. 1974. The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association*, 47:5-22.

Rosati, Connie S. 2000. Brandt's Notion of Therapeutic Agency. *Ethics*, 110: 780-811.

Russell, Bruce. *A Priori* Justification and Knowledge. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/apriori/

Sabini, John and Maury Silver. 1982. *Moralities of Everyday Life*. Oxford: Oxford University Press.

Sanfrey, A.G., J.K. Rilling., J.A. Aronson, L.E. Nystrom, L.E., and J.D. Cohen. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300: 1755-1758.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, Massachusetts: Harvard University Press.

Scanlon, T.M. 2000. Intention and Permissibility. *Proceedings of the Aristotelian Society*, 74 (Suppl): 301-317.

Skorupski, John. 1999. *Ethical Explorations*. New York: Oxford University Press.

Smith, Malcolm B.E. 1977. Rawls and Intuitionism. *Canadian Journal of Philosophy*, Supplementary Volume 3: 163-178.

Smith, Malcolm B.E. 1979. Ethical Intuitionism and Naturalism: A Reconciliation. *Canadian Journal of Philosophy*, 9: 609-629.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell Publishing Ltd.

Singer, Peter. 1972. Famine, Affluence, and Morality. *Philosophy and Public Affairs*, 1: 229-243

Singer, Peter. 1974. Sidgwick and Reflective Equilibrium. *Monist*, 58: 490-517.

Singer, Peter. 1975. *Animal Liberation*. New York: Harper Collins Books. 2nd edition, New York Review/Random House, 1990.

Stampe, D. 1977. Toward a Causal Theory of Linguistic Representation. in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds) *Midwest Studies in Philosophy: Studies in the Philosophy of Language*, vol. 2. Minneapolis: University of Minnesota Press, 81-102.

Stalnaker, Robert C. 1984. *Inquiry*. Cambridge: MIT Press.

Stanley, Jason and Timothy Williamson. 1995. Quantifiers and Context Dependence. *Analysis*, 55: 291-295.

Stanley, Jason and Zoltan G. Szabo. 2000. On Quantifier Domain Restriction. *Mind and Language*, 15: 219-261.

Stevenson, Charles L. 1937. The Emotive Meaning of Ethical Terms. *Mind*, 46:14-31.

Stevenson, Charles L. 1944. *Ethics and Language*. New Haven and London: Yale University Press.

Strawson, P.F. 1968. Freedom and Resentment. In *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies*, 127(1): 109-166.

Taurek, John M. 1977. Should the Numbers Count? *Philosophy and Public Affairs*, 6: 293-316.

Taylor, Gabriele. 1985. *Pride, Shame, and Guilt: Emotions of Self-Assessment*. Oxford: Oxford University Press.

Thomson, Judith Jarvis. 1985. "The Trolley Problem." *Yale Law Journal*, 94: 1395-1415. Reprinted in J.J. Thomson and William Parent, eds., *Rights, Restitution, and Risk: Essays in Moral Theory*. Cambridge: Harvard University Press, 1986, p.94-116.

Tooley, Michael. 1980. An Irrelevant Consideration: Killing Versus Letting Die. In B. Steinbock and A. Norcross (eds) *Killing and Letting Die*, New York:Fordham University Press.

Tresan, Jon. 2006. *De Dicto Internalist Cognitivism*. *Nous*, 40: 143-165.

Trivers, R.L. 1971. *The Evolution of Reciprocal Altruism*. Quarterly Review of Biology, 46: 35-57.

Tye, M. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, Massachusetts: MIT Press.

Tye, M. 2003. A Theory of Phenomenal Concepts. In Anthony O'Hear (ed) *Minds and Persons: Royal Institute of Philosophy Supplement: 53*. Cambridge: Cambridge University Press, 91-105.

Ullmann-Margalit, Edna, and Sidney Morgenbesser. 1977. Picking and Choosing. Social Research, 44: 757-785.

Unger, Peter. 1996. *Living High and Letting Die*. New York: Oxford University Press.

Velleman, J. David. 1988. Brandt's Definition of "Good." *The Philosophical Review,* 97: 353-371.

Velleman J. David. 1992. The Guise of the Good. *Nous,* 26: 3-26.

Velleman, J. David. 2000. *The Possibility of Practical Reason*. New York: Oxford University Press.

Velleman, J. David. 2002. Motivation by Ideal. *Philosophical Explorations,* 5: 90-104.

Wedgewood, Ralph. 2004. The Metaethicists' Mistake. *Philosophical Perspectives,* 18: 405-426.

Williams, Bernard. 1976. "Persons, character, and morality." In A.O. Rorty (ed.), *The identities of persons*, Berkeley: Universtiy of California Press. Reprinted in B. Williams, *Moral Luck,* Cambridge: Cambridge University Press, 1981.

Williams, Bernard. 1981. Internal and External Reasons. In B. Williams, *Moral Luck*, Cambridge: Cambridge University Press. Reprinted in n S. Darwall, A. Gibbard, and P. Railton, eds., *Moral Discourse and Practice,* New York: Oxford University Press, 1997.

Zimbardo, Philip G. and Ann L. Weber. 1997. *Psychology*. 2nd Edition. New York: Longman.

Zimbardo, Philip G. 2007. *The Lucifer Effect: Understanding How Good People Turn Evil*. New York : Random House.